Zur Evaluation von Gewaltpräventionsmassnahmen : Drei Analysen zur Wirksamkeit von Interventionen

Eisner, Manuel ; Ribeaud, Denis

Universität Zürich
Pädagogisches Institut

University of Cambridge
Institute of Criminology

z-proso *Zürcher Projekt zur sozialen Entwicklung von Kindern*

ZUR EVALUATION VON GEWALTPRÄVENTIONSMASSNAHMEN

DREI ANALYSEN ZUR WIRKSAMKEIT VON INTERVENTIONEN

Manuel Eisner, Denis Ribeaud

Herausgeber: Manuel Eisner und Denis Ribeaud

Forschungsbericht aus der Reihe *z-proso*
Zürcher Projekt zur sozialen Entwicklung von Kindern

Zürich, Juli 2007, Bericht Nr. 06

Inhaltsverzeichnis

Markt, Macht und Wissenschaft;
Kritische Überlegungen zur deutschen Präventionsforschung

Wer sich auf manchen deutschsprachigen Internetseiten zur Gewaltprävention umsieht, kann den Eindruck gewinnen, heutige Präventionsprogramme seien geradezu Wunderwerke wirksamer Sozialtechnologie. Lehrer schreiben begeistert, wie phantastisch das Klassenklima nach Umsetzung eines Sozialkompetenzprogramms ist; Eltern werden zitiert, wie sich ihr verhaltensauffälliger Racker dank Erziehungskursen in ein Musterkind verwandelt hat; und man schmückt sich mit Presseberichten, in denen erklärt wird, ab sofort würden dank des neuen Programms die Kinder abgeklärt miteinander reden statt sich zu verprügeln. Selbstverständlich sagt solche als „Evaluation" kaschierte Eigenwerbung über die tatsächliche Wirksamkeit nicht mehr aus als begeisterte Lesermeinungen zu den Effekten von Wünschelruten, Kupferbändern oder Kristallkugeln, nämlich gar nichts.

Das gleiche gilt für reine Prozessevaluationen - also Einschätzungen darüber, wie das Projekt umgesetzt wurde und wie zufrieden die Benutzer mit dem Programm sind. Zwar wird gerade gegenüber der Praxis nicht selten der Eindruck geweckt, dass glückliche Projektteilnehmer ein Gradmesser für ein gutes Programm seien, oder dass man Wirksamkeit bestimmen könne indem man frage, ob die Umsetzenden das Programm für wirksam halten. Tatsächlich sind weder zufriedene Teilnehmer noch von einer Wirkung subjektiv überzeugte Umsetzende ein wissenschaftlich annehmbares Kriterium für die Frage, ob tatsächlich die gewünschten Änderungen erreicht wurden.

Vielmehr besteht unter Präventionsforschenden weit herum Einigkeit darüber, dass für einen wissenschaftlich abgestützten Nachweis von positiven Wirkungen höhere Ansprüche an die Forschungsanlage gestellt werden müssen. Worin diese Anforderungen bestehen, ist dank der Pionierarbeiten von Cochrane (1972), Campbell und Stanley (1966), Cook und Campbell (1979), sowie Shadish et al. (2002) weitgehend geklärt. Es handelt sich im Wesentlichen um methodisch sorgfältig angelegte Experimentalstudien, bei denen idealerweise die Teilnehmenden zufällig aus der Zielpopulation ausgewählt werden und die Zuweisung zu den Behandlungsbedingungen nach dem Zufallsprinzip erfolgt. Durch den Vergleich von Veränderungen zwischen Kontroll- und Interventionsgruppe erlauben sie, sofern keine Annahmen von Experimentaldesigns verletzt werden, Aussagen über die erzielten Wirkungen mit einer hohen Gültigkeit.

Um diese Erkenntnis herum ist in den vergangenen 20 Jahren im angelsächsischen Sprachraum eine wissenschaftliche Bewegung entstanden, welche sich dem Ziel einer *evidenzbasierten Prävention* verschrieben hat. Sie kann als das Bestreben definiert werden, Fehlschlüsse über die Wirkungen von Präventionsmaßnahmen zu vermeiden und Maßnahmen zur Verminderung oder Verhinderung von unerwünschten Verhaltensweisen möglichst weitgehend auf gesichertes empirisches Wissen abzustützen (für den Bereich der Kriminalprävention vgl. Rössner, Bannenberg, & Landeshauptstadt Düsseldorf, 2002; Sherman, Farrington, Welsh, & MacKenzie, 2002). Evidenzbasierte Prävention beruht auf dem Grundsatz, dass die Wirksamkeit von Prävention durch gute empirische Forschung überprüft werden kann und dass durch den Zusammenzug der Forschungsergebnisse zuverlässige Kenntnisse darüber gewonnen werden können, welche Präventionsmaßnahmen wirksam, wirkungslos oder schädlich sind; wie Maßnahmen, welche sich in der Forschung als wirksam erwiesen haben, effektiv in die Praxis umgesetzt werden können; wie sich Maßnahmen an die Bedürfnisse unterschiedlicher Bevölkerungsgruppen anpassen lassen: und welche Aspekte der praktischen Umsetzung dafür verantwortlich sind, dass positive Wirkungen erzielt werden können.

Infolge des gestiegenen Interesses an besserer, evidenzgestützter Gewalt- und Kriminalprävention ist es in den letzten 10 Jahren zu einer deutlichen Zunahme von wissenschaftlichen Wirkungsevaluationen mit einem randomisierten Kontrollgruppendesign gekommen. Im Bereich der entwicklungsorientierten Gewaltprävention gehören hierzu im deutschsprachigen Raum beispielsweise die Studien von Heinrichs et al. (2006) über *Triple P*, von Schick und Cierpka (2005; 2006) zu *Faustlos*, von Asshauer und Hanewinkel (2000) über das Lebenskompetenztraining *Fit und Stark fürs Leben*, von Lösel et al. (2006) zu einem kombinierten Elterntraining und Sozialkompetenzprogramm *EFFEKT* sowie von Eisner et al. (2007) über eine weitere Kombination von

Elterntraining und schulischen Sozialkompetenztraining. Mehrere größere Studien sind gegenwärtig im Gang oder haben noch keine Evaluationsergebnisse publiziert.

Es gibt zudem gegenwärtig in Politik und Praxis eine erfreulich große Bereitschaft, den Argumenten für eine evidenzbasierte Präventionspolitik Folge zu leisten und die damit einhergehenden zeitlichen und finanziellen Kosten auf sich zu nehmen. Allerdings kann diese Bereitschaft nur aufrecht gehalten bleiben, wenn die Wissenschaft ihre Funktionen so gut als möglich wahrnimmt. Hierzu gehört, dass sie Forschung so durchführt, dass die Ergebnisse unverzerrt die tatsächlichen Wirkungen eines Programms oder einer Maßnahme wiedergeben und dass sie die Ergebnis möglichst so kommuniziert, dass Außenstehende Dritte ein angemessenes Bild der Sachlage gewinnen.

Nun werden viele dieser Studien zur Wirkung von Präventionsmaßnahmen von Forschergruppen durchgeführt, welche das Produkt selbst entwickelt oder in Lizenz übernommen haben und es kommerziell vertreiben. Man kann daher auch von *Eigenevaluationen* sprechen. Dem stehen Fremdevaluation gegenüber, bei welchen die Durchführung der Studie, die Analyse der Daten sowie die Publikation der Ergebnisse und Folgerungen von Forschenden durchgeführt werden, die kein direktes Eigeninteresse am evaluierten Programm haben.

Es ist grundsätzlich zu begrüßen, wenn Programmentwickler und -vertreiber ihre Maßnahmen einer wissenschaftlichen Evaluation unterziehen. Dass hierbei Eigeninteressen bestehen, ist an sich nicht problematisch, sofern ausreichend wirksame Kontrollmechanismen bestehen, um die Qualität der Forschung zu garantieren. Es besteht aber kein Zweifel, dass diese Eigeninteressen nicht identisch sind mit dem Interesse von Öffentlichkeit und Fachpublikum, möglichst unverzerrte und wirklichkeitsnahe Schätzwerte der tatsächlichen Effekte eines Präventionsprogramms zu erhalten. Und es wäre naiv zu glauben, dass dieser Interessenkonflikt in jedem Fall in eine unverzerrte Darstellung der erzielten Wirkungen mündet.

Dies dokumentiert eine wichtige Studie von Petrosino und Soydan (2005). Die Autoren haben im Rahmen einer Metaanalyse untersucht, wie stark die publizierten Wirkungen einer Evaluation davon abhängen, ob die Studie durch Programmentwickler und Programmvertreiber oder durch unabhängige Wissenschaftler durchgeführt wurde. Sie haben zu diesem Zweck 300 randomisierte Studien untersucht, welche die Wirkungen von Interventionsprogrammen auf die Rückfallwahrscheinlichkeit von Straftätern prüften. Alle Studien wurden anschließend bezüglich der Verbindung zwischen Programmentwicklern und Projektleitung bewertet. Die Auswertung zeigte für 137 Eigenevaluationen einen schwachen, aber beachtlichen positiver Effekt von Cohen's d = 0.16. Dem steht gegenüber, dass in den 124 Fremdevaluationen die durchschnittliche Effektstärke ziemlich genau bei Null lag (Cohens d = 0.02). Mit anderen Worten: Während man im Durchschnitt von Eigenevaluationen zum Schluss gelangt, dass heutige Interventionsprogramme zur Reduktion der Rückfallwahrscheinlichkeit eine gewisse Wirkung haben, führt eine Betrachtung der Fremdevaluationen zum Schluss, dass die Programme völlig wirkungslos sind.

Die Gründe für die Diskrepanz zwischen Eigenevaluationen und Fremdevaluationen werden unterschiedlich beurteilt. Petrosino und Soydan (2005) unterscheiden zwei Interpretation, die sie als Umsetzungsperspektive (*high fidelity view*) und als zynische Perspektive (*cynical view*) bezeichnen. Die positive Interpretation lautet, dass eine aktive Beteiligung der Programmvertreiber dazu führt, dass die Programme in besonders guter Qualität, mit Enthusiasmus und in großer Umsetzungstreue realisiert werden – und sich dies dann in einer positiven Wirkung niederschlägt. Das einzige Problem wäre dann, wie man all diese Komponenten eines idealen Modellversuchs auch in Bedingungen realisieren kann, unter denen die Programmentwickler nicht persönlich anwesend sind – also den Regelfall von Präventionsprojekten.

Die zynische Perspektive hingegen argumentiert, dass eine Reihe von subtilen und weniger subtilen Faktoren die Programmentwickler und –vertreiber unter Druck setzen, positive Ergebnisse zu produzieren. Dieses Eigeninteresse erstreckt sich auf die *wissenschaftliche Reputation*, da man mit einem Programm meist die eigenen theoretischen Annahmen prüft; auf die *zeitlichen Investitionen*, da Entwicklung und Test eines Programms bis zur Marktreife oft mehrere Jahre in Anspruch nehmen; auf *Forschungsgelder*, da positive Ergebnisse mit einer größeren Wahrscheinlichkeit weiterer Unterstützung einher gehen; auf *politischen Einfluss*, da Vertreiber von erfolgreichen Programmen in Verwaltung und Politik mehr Gehör finden; sowie direkte

finanzielle Eigeninteressen, da für Entwicklung und Vertrieb eines Programms oft erhebliche Investitionen notwendig sind, die nur bei positiven Testresultaten wieder eingeholt werden können.

Ein verzerrtes Bild der Wirkungen von Präventionsprogrammen ist aus mehreren Gründen problematisch: Erstens wird eine optimistische Einschätzung von Interventionswirkungen in der Regel dazu führen, dass politische Entscheidungsträger zusätzliche öffentliche Mittel in die entsprechende Maßnahme investieren. Wenn allerdings diese optimistische Einschätzung auf verzerrten Ergebnissen basiert, dann bedeutet das eine Fehlallokation von öffentlichen Ressourcen. Dies führt zu einem Entzug von Mitteln für andere Programme, die möglicherweise in Wirklichkeit ebenso wirksam oder gar wirksamer sind.

Zweitens leiten systematisch verzerrte Ergebnisse von Einzelstudien die wissenschaftliche Forschung auf Abwege und erschweren den Fortschritt in der Entwicklung besserer Präventionsmaßnahmen. Ein Beispiel sind Metaanalysen, in denen die Ergebnisse vieler Einzelstudien zusammengefasst und in statistischen Kennwerten ausgedrückt werden. Metaanalyse gilt als der beste Weg, ein von subjektiven Eindrücken unverzerrtes Bild des Wissensstandes zu einem Interventions- oder Präventionsbereich zu erhalten. Dies gilt aber nur, wenn die Einzelstudien selbst als unverzerrte Schätzungen der tatsächlichen Wirkungen einer Massnahme gelten können. Wenn in eine Metaanalyse mehrere, systematisch in die gleiche (meistens zu optimistische) Richtung verzerrte Einzelstudien eingehen, dann führen sie zu einer irrtümlichen Bilanz des aktuellen Forschungsstandes.

Drittens schließlich bewirken systematisch verzerrte Ergebnisse von Evaluationsforschungen einen Vertrauensverlust der Öffentlichkeit in die wissenschaftliche Wirkungsevaluation. Beispielsweise wird es auf die Dauer außenstehenden Beobachtern nicht verborgen bleiben, wenn die Effekte von Evaluationsstudien immer das vom Studienleiter vertretene Präventionsprogramm stützen oder wenn sich nach einer Weile herausstellt, dass die Umsetzung eines Programms im Alltag nicht die wissenschaftlich versprochenen Wirkungen erbringt. Hierdurch kann es geschehen, dass sich öffentliche Hand und Wissenschaftsförderung enttäuscht von den meist kostspieligen experimentellen Evaluationsprojekten abwenden, da sie das Versprechen einer unverzerrten Beurteilung von Interventionswirkungen nicht einzulösen vermögen.

Empirische Illustrationen

Der folgende Abschnitt diskutiert an ausgewählten Beispielen drei Probleme, die mit Eigenevaluationen einher gehen können: Die Schwierigkeit, Befunde aus Eigenevaluationen zur Wirksamkeit von Präventionsprogrammen in Fremdevaluationen zu replizieren; das Problem, dass im Forschungsprozess mehrere methodische Entscheide so gefällt werden, dass sie das gewünschte Ergebnis begünstigen und die Befunde daher nicht valide sind; und die Frage, ob Forschungsbefunde gegenüber Praxis und Öffentlichkeit fair dargestellt werden.

Neben Beispielen aus der US-amerikanischen Forschung werden hierbei auch Beispiele aus der deutschsprachigen Forschung aufgeführt. Die Auswahl reflektiert die Interessen und Kompetenzen des Autors. Sie soll nicht als Stellungsnahme für oder gegen die diskutierten Programme interpretiert werden, sondern dient einzig der Illustration von Konflikten, die in der Struktur von Eigenevaluationen angelegt sind.

Ergebnisse aus Eigenevaluationen können in Fremdevaluationen nicht repliziert werden

Experimentelle Wirkungsevaluationen werden in der Regel durchgeführt, um aufgrund der gemessenen Effekte Aussagen über die zu erwartenden Wirkungen auch außerhalb der Untersuchung machen zu können. Man spricht hier von *externer Validität* (vgl. z.B. Shadish et al., 2002). Damit aber diese Folgerung zulässig ist, müssen die gemessenen Effekte *unverzerrte Schätzwerte* der tatsächlich erzielbaren Wirkungen sein – sie sollen also die effektive Wirkung weder systematisch über- noch unterschätzen. Nur dann gilt, dass die Studienergebnisse auch auf die Welt außerhalb der besonderen Bedingungen einer experimentellen Studie generalisiert werden können. Allerdings scheint dies häufig nicht zuzutreffen, wie zwei Beispiele illustrieren.

Beim ersten Beispiel handelt es sich um das Suchtpräventionsprogramm *ALERT*. ALERT ist ein schulbasiertes Programm zur Förderung von Lebenskompetenzen, das in den USA weit verbreitet ist und als wissenschaftlich

gut abgestützt gilt (vgl. die Webseite www.projectalert.best.org). Es besteht aus insgesamt 14 Lektionen im siebten und achten Schuljahr. Gemäß der Webseite der Betreiber wird das Programm auf sieben Empfehlungslisten von US-amerikanischen Behörden - darunter dem Bildungsministerium, dem Gesundheitsministerium und dem Justizministerium - als evidenzbasiertes Modellprogramm empfohlen. Die Vertreiber von ALERT werben damit, dass das Programm beispielsweise den Marijuana-Gebrauch um 60 Prozent reduziert, den Nikotinmissbrauch um 35-55 Prozent vermindert und den Alkoholmissbrauch signifikant senkt (Ellickson, 1998; Ellickson & Bell, 1990; Ellickson, Bell, & Harrison, 1993; Ellickson, Bell, & McGuigan, 1993).

Aufgrund diese positiven Erfolgsmeldungen unterwarfen St. Pierre et al (2006) das Programm erstmals einer unabhängigen Evaluation. Mit über 1600 Schülern und einem randomisierten Kontrollgruppendesign handelte es sich um eine sorgfältig angelegte Längsschnittstudie, die höchsten Qualitätsanforderungen an den Wirkungsnachweis von Interventionen genügt.

Die Ergebnisse zeigten für *keine* der Variablen zur Messung von Veränderungen des Substanzkonsums einen positiven Effekt. Hingegen finden sie einen möglichen negativen (d.h. unerwünschten) Effekt des Programms auf den Substanzkonsum, wobei sie aber darauf hinweisen, dass dieser angesichts der vielen gemessenen Zielgrößen möglicherweise zufällig zustande gekommen ist. Ebenso wenig konnte eine systematische Wirkung auf die Mediatoren (d.h. direkt durch das Programm angesprochenen Zielgrößen wie beispielsweise Einstellungen zu Drogen) gezeigt werden; die wenigen Effekte verteilten sich gleichermaßen auf positive und negative Wirkungen.

Ein analoges Problem kann im deutschsprachigen Raum für *Triple P* beobachtet werden. Triple P ist ein vom Matthew Sanders an der Queensland Universität in Australien entwickeltes Elterntraining. Es fußt auf verhaltenstheoretischen Grundlagen und will dem Problemverhalten von Kindern und Jugendlichen vorbeugen (Sanders, 1999; Sanders, Lynch, & Markie-Dadds, 1994). Die internationale Webseite von Triple P (www.triplep.net) wirbt mit dem unbescheidenen Slogan „parenting now comes with an instruction manual". Das Programm wurde seit seiner Entwicklung in über 30 Studien, von denen allerdings viele auf einer sehr kleinen Zahl von Teilnehmern basieren, auf seine Wirksamkeit hin überprüft. Dabei kommen ausnahmslos alle publizierten Studien der Programmentwickler zum Befund, dass Triple P gute Wirkungen auf das Erziehungsverhalten und das Problemverhalten der Kinder zeige. Es findet sich daher auch auf Empfehlungslisten etwa der WHO und des Europarates.

Seit 2001 wird Triple P auch in Deutschland und der Schweiz vertrieben. Im Jahr 2006, mehrere Jahre nachdem der Vertrieb bereits angelaufen war, publizierten die Vertreiber der deutschen Version Ergebnisse ihrer Evaluationsstudie (Heinrichs et al., 2006). In dieser Studie wurde Triple P als universelle Maßnahme den Eltern von Kindergärtnern in Braunschweig angeboten. Die Studie berichtet von durchwegs positiven Effekten. Insbesondere kommt sie zum Schluss, dass durch Triple P das Problemverhalten der Kinder signifikant zurückgegangen sei. Die Forschergruppe empfahl daher in den Folgerungen „eine breitflächige Dissemination dieses Trainings" (Heinrichs et al., 2006: 94). Triple P wurde im Rahmen des Zürcher Projektes zum sozialen Verhalten von Kindern (z-proso) erstmals einer unabhängigen Evaluation unterzogen (Eisner et al., 2007). Die Studie ist mit rund 1300 teilnehmenden Kindern und Eltern weltweit eines der größten Projekte zur Evaluation der Wirkung von Elterntrainings. Das Studiendesign wurde in Zusammenarbeit mit renommierten Experten der Evaluationsforschung als Längsschnittstudie angelegt, in die ein randomisiertes Kontrollgruppendesign eingebettet ist. Das Verhalten der Kinder wurde in einem Mehrinformandenansatz aus der Perspektive der Eltern, der Lehrpersonen sowie der Kinder selbst untersucht. Der Kurs selbst wurde von erfahrenen und gemeinsam mit Triple P Schweiz ausgewählten Trainerinnen vermittelt. Rund 31 Prozent der Familien, welche in der Interventionsgruppe an der Längsschnittstudie teilnehmen, absolvierten den Kurs. Die Zufriedenheit unter den Kursteilnehmern war hoch.

Die Teilnahmerate und die Zufriedenheit der Eltern mit dem Kurs entsprechen weitgehend jenen der Braunschweiger Studie. Dies gilt allerdings nicht für die Ergebnisse zu den Wirkungen: Eine so genannte Intention-to-treat Analyse ergab weder für die sieben Teilaspekte von elterlichem Erziehungsverhalten noch für die neun Teilindikatoren von kindlichem Problemverhalten irgendwelche signifikanten positiven Effekte. Beschränkt man die Analyse nur auf jene Eltern, welche tatsächlich den größten Teil des Programms absolviert hatten, dann ergaben sich einige schwach positive Effekte auf das Erziehungsverhalten. Dem steht gegenüber,

dass ein statistisch signifikanter negativer (d.h. unerwünschter) Effekt für nicht-aggressives externalisierendes Problemverhalten des Kindes auf der Perspektive der Lehrperson gefunden wurde.

Für beide Beispiele ergibt sich: Studien, welche als Eigenevaluationen durchgeführt werden, berichten positive Wirkungen. Unabhängige Studien hingegen finden nichts. Solche Ergebnisse sind in der Forschung zu evidenzbasierter Prävention keine Einzelfälle. Ähnliche Ergebnisse werden auch in anderen Präventionsbereichen berichtet. In der Metaanalyse von 84 Studien zur Wirkung von *Sozialkompetenzprogrammen* durch Lösel und Beelmann (2003) beispielsweise lag die mittlere Effektstärke bei Eigenevaluationen mit Cohens d=0.49 fast doppelt so hoch wie bei jenen Programmen, die von unabhängigen Forschenden realisiert wurden. Eine Metaanalyse von Borman et al. (2003: 37) für *Schulreformprogramme* fand, dass "studies performed by the developer yielded considerably stronger effects than studies performed by others". Und eine kürzlich publizierte Studie, welche die fünf meistempfohlenen *Drogenpräventionsprogramme* der USA unter die Lupe genommen hat (Gandhi, Murphy-Graham, Petrosino, Chrismer, & Weiss, 2007) kommt zum Schluss, dass die publizierten Ergebnisse zur Wirksamkeit der Programme selektiv nur die besten Befunde rauspicken, dass unabhängige Evaluationen weitgehend fehlen und dass dort, wo Replikationen realisiert wurden, die Wirkungen oft nicht bestätigt werden konnten.

Positive Ergebnisse können das Ergebnis einer Kette von methodisch problematischen Entscheiden sein

Es kann mehrere Gründe dafür geben, weshalb Ergebnisse von Selbstevaluationen in Fremdevaluationen nicht repliziert werden können. Dabei kann sicher eine Rolle spielen, dass die Umsetzungsqualität in Fremdevaluationen nicht in gleichem Masse garantiert ist wie bei Eigenevaluationen. Allerdings gibt es aus verschiedenen Studien auch Hinweise darauf, dass die positiven Ergebnisse von Selbstevaluationen damit zusammenhängen können, dass in jeder Evaluationsstudie eine Vielzahl von aufeinander folgenden methodischen Entscheiden getroffen werden muss. Wenn jeder dieser Entscheide nur geringfügig durch das gewünschte Ergebnis eingefärbt ist, kann hieraus im Endergebnis ein problematisches Gesamtbild entstehen.

Eine systematische Analyse von Littell (2005) hat kürzlich das Problem in grellem Licht aufscheinen lassen. Sie untersuchte im Auftrag der *Campbell Collaboration* systematisch den Wissensstand bezüglich der Wirkungen von multisystemischer Therapie (MST) bei der Behandlung von verhaltensauffälligen und delinquenten Jugendlichen (Littell, 2005). Multisystemische Therapie (Henggeler et al., 1996; Henggeler, Melton, & Smith, 1992) gilt als eines der erfolgreichsten Programme zur Reduktion von Verhaltensproblemen bei Hochrisikopopulationen. Es steht als Modellprogramm auf vielen Empfehlungslisten US amerikanischer Behörden und wird in den USA und in Europa bei jährlich rund 10,000 jugendlichen Straftätern eingesetzt. Anfangs der 2000er Jahre erhielt die Gruppe um die Programmentwickler jährlich über 20 Millionen US-Dollar an Forschungsaufträgen.

Dieser Erfolg gründet fast ausschließlich auf wissenschaftlichen Publikationen der Programmentwickler, welche durchwegs über positive Effekte von MST berichten (für eine Übersicht aller Publikationen zu Experimentalstudien mit MST vgl. Littell, Popa, & Forsythe, 2005). Die systematische Review von Littell (2005) nahm diese Effekte genauer unter die Lupe, indem sie neben den publizierten Studien der Programmentwickler auch unpublizierte Studien und unabhängige Studien berücksichtigte. Zudem zog sie alle empirisch getesteten (und nicht nur die positiven) Effektgrößen in Betracht und unterzog die Verfahren bei der Zuordnung zu den Behandlungsgruppen einer genauen Prüfung.

Sie kam zum Schluss, dass zwar keine Hinweise auf eine schädliche Wirkung von MST bestünden, dass aber die bisherigen Ergebnisse keine Unterstützung für die Hypothese liefern, MST sei wirksamer als herkömmliche Therapieverfahren. Die sehr viel positivere Selbstdarstellung von MST erklärt sie als Folge einer Kombination von systematischen Verzerrungen. Hierzu gehören: Studien mit unerwünschten Ergebnissen wurden nicht publiziert und verzerren daher den Eindruck bei den publizierten Arbeiten; in gewissen Studien mit positiven Ergebnissen ist das Vorgehen bei der Zuteilung zu Interventions- und Experimentalgruppe unklar; zwischen den anfänglichen unveröffentlichten wissenschaftlichen Schlussberichten und den publizierten Ergebnissen bestehen ungeklärte

Diskrepanzen, welche sich zu Gunsten der Behandlung auswirken; was als Behandlungsabschluss gilt, wurde subjektiv definiert und ist damit möglichen unbewussten Manipulationen ausgesetzt.[1]

Vor diesem Hintergrund lohnt es sich, die oben skizzierte Diskrepanz zwischen den sehr positiven Ergebnissen der Eigenevaluation von Triple P in Braunschweig (Heinrichs et al., 2006) und den deutlich weniger erfreulichen Befunden der Fremdevaluation von Eisner et al. (2008) nochmals unter die Lupe zu nehmen. Eine solche Betrachtung ergibt, dass analog zu den Befunden von Littell (2005) in der Studie von Heinrichs et al. (2006) mehrere methodisch problematische Entscheide gefällt wurden, welche allesamt zum Effekt haben, dass die Interventionseffekte überschätzt werden (vgl. auch Eisner & Ribeaud, 2008). Sie seien hier kurz zusammengefasst:

Erstens haben in Braunschweig knapp 70 Prozent der angesprochenen Eltern die Teilnahme an der Studie völlig verweigert, so dass über diese Eltern keinerlei Daten vorliegen. Es handelt sich hierbei überwiegend um bildungsferne und sozial unterprivilegierte Eltern. Eine solch tiefe Teilnahmerate bewirkt, dass eine inferenzstatistische Absicherung (d.h. der Schluss von den erhobenen Daten auf die Grundbevölkerung) der Befunde kaum möglich ist.

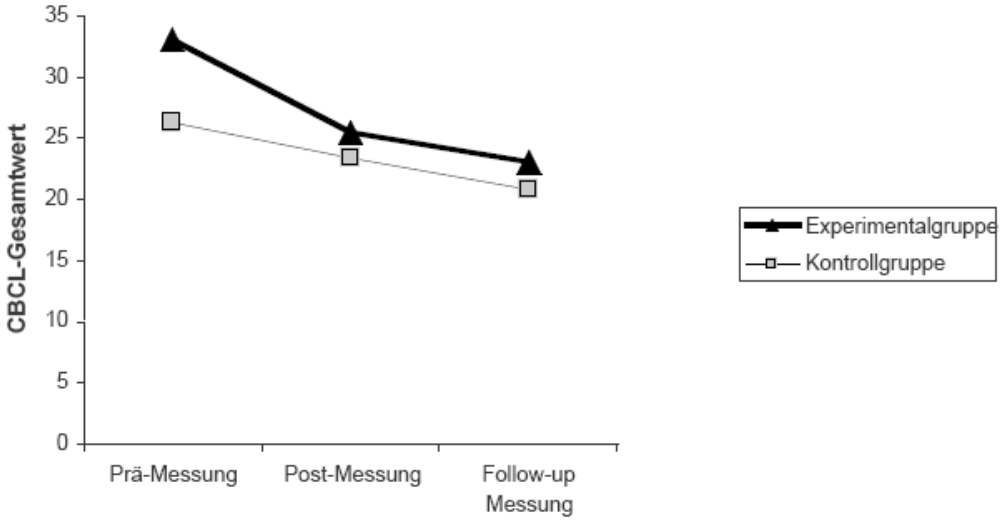
Vielmehr ähneln Rückschlüsse auf die Gesamtpopulation, aus der die Stichprobe gezogen wurde, eher einer Lotterie als einer wissenschaftlichen Analyse. Dies gilt besonders, wenn beansprucht wird, eine so genannte Intention-to-Treat Analyse durchzuführen. Solche Analysen sind immer Analysen *aller* Personen, welche einer Behandlungsbedingung zugewiesen wurden. Wenn für sieben von zehn potentiellen Teilnehmern aber keine Daten vorliegen, dann ist auch eine entsprechende Analyse unmöglich.

Zweitens wurden Projektteilnehmer für die Datenanalyse nachträglich aus der Behandlungsbedingung in die Kontrollbedingung umgeteilt. So heißt es in Heinrichs et al (2006: 86), dass „Familien, die zur Kontrollbedingung randomisiert wurden (N = 62) und Triple P Ablehner (N = 28) [….] für die folgenden Auswertungen zusammengefasst wurden". Mit anderen Worten: Jene Familien, die für eine Teilnahme an Triple P vorgesehen waren aber das Programm nicht nutzen wollten, wurden im Nachhinein der Kontrollgruppe zugewiesen. Diese Umteilung ist eine grobe Verletzung der methodischen Prinzipien von randomisierten Forschungsdesigns. Es ist unverständlich, wie die Autoren ihre Studie dennoch als randomisiertes Kontrollgruppendesign darstellen können.

Drittens unterscheiden sich die Ausgangswerte (d.h. vor der Intervention) zwischen Kontrollgruppe und Behandlungsgruppe bei mehreren zentralen Zielvariablen massiv. Dies verletzt eine zentrale Annahme von randomisierten Experimentalstudien, nämlich dass die verglichenen Gruppen hinsichtlich der Kriteriumswerte vor der Intervention nicht unterscheiden. Die Auswirkungen der Verletzung dieser methodischen Anforderung sind in Abbildung 1 illustriert, welche den Verlauf des Gesamtwertes von externalisierendem Problemverhalten in der Kontrollgruppe und der Interventionsgruppe zwischen der Prä-Messung, der Post-Messung und der Follow-up Messung zeigen. Die Daten sind der Publikation von Heinrichs et al. (2006: 89) entnommen.

[1] Seit der Publikation der Kritik von Littell sind mehrere neue Arbeiten zu multisystemischer Therapie erschienen, die allerdings ein eher positives Bild der Wirkungen zeichnen. Unabhängige Evaluationen von Timmons-Mitchell et al. (2006) in den USA sowie von Ogden und Hagen (2006) in Norwegen fanden positive Effekte; die ebenfalls unabhängige Evaluation von Olsson und Sundell (2008) in Schweden hingegen fand keine Effekte.

Abbildung 1 Ein Beleg für die Wirksamkeit von Triple P? Mittelwerte von kindlichem Problemverhalten in Experimental- und Kontrollgruppe gemäß Heinrichs et al. (2006: 89)



Hinweis: Mittelwerte der Gesamtscores für das Problemverhalten des Kindes (Child Behavior Checklist) nach Einschätzung der Mutter. Die Originaldaten wurden aus Heinrichs et al. (2006: 89) übernommen.

Die Autoren interpretieren diesen Datenverlauf als Beleg für die Wirksamkeit von Triple P, da der Rückgang in der Triple P Gruppe größer ist als in der Kontrollgruppe. Dem steht gegenüber, dass ein Datenverlauf, wie er in Abbildung 1 gezeigt ist, geradezu als Textbuchbeispiel für einen Regressionseffekt und damit ein Methodenartefakt interpretiert werden kann (für eine Diskussion des Problems vgl. Nachtigall & Suhl, 2002: 7). Denn er zeigt im Wesentlichen, dass sich Gruppen mit hohen anfänglichen Unterschieden im Verlauf der Zeit einander angleichen. Wichtig ist in diesem Zusammenhang, dass die Werte der Triple P Gruppe zu keinem Messzeitpunkt nach der Intervention besser werden als jene in der Kontrollgruppe.

Die hier geschilderten Probleme unterstellen nicht, dass die Daten absichtlich manipuliert wurden, um positive Ergebnisse zu erzielen. Sie dokumentieren aber, dass die von Petrosino und Soydan (2005) „zynisch" genannte Interpretation der Diskrepanzen zwischen Eigenevaluationen und Fremdevaluationen nicht leichtfertig vom Tisch gefegt werden sollte: Vielmehr wäre es erstaunlich, wenn es bei Eigenevaluationen nie zu Situationen käme, in denen Evaluatorenteams mit einem Eigeninteresse an den Ergebnissen Gefahr laufen, problematische methodische Entscheide und Interpretationen so miteinander zu verknüpfen, dass die publizierten Ergebnisse nicht als unverfälschte Schätzungen der tatsächlich erzielbaren Wirkungen interpretiert werden können.

Diskrepanzen zwischen der öffentlichen Darstellung und den tatsächlich nachweisbaren Effekten von Programmen

Drittens soll ein Problem aufgegriffen werden, das in der bisherigen Fachdiskussion zur Qualitätssicherung von Evaluationsstudien nur am Rande gewürdigt wurde. Es bezieht sich auf die Frage, ob die Ergebnisse von Eigenevaluationen *gegenüber der Öffentlichkeit* fair dargestellt werden.

Dabei müssen zwei Interessen gegeneinander abgewogen werden: So ist es einerseits völlig legitim, dass Programmbetreiber auf Webseiten und in Prospekten die Vorteile ihrer Produkte herausstreichen, positive Ergebnisse von wissenschaftlichen Studien betonen und die gemessenen Wirkungen in eine einfache Sprache übersetzen, soweit sich die Darstellung im Rahmen einer populären Interpretation der tatsächlich erzielten Effekte bewegt. Dem steht entgegen, dass Praxis und Laienpublikum ein Interesse daran haben, ein wirklichkeitsnahes Bild des Wissensstandes zu erhalten und zuverlässig über das Wirkungspotential eines Präventionsprogramms unterrichtet zu werden.

Das Problem soll am Beispiel des Gewaltpräventionsprogramms *Faustlos* illustriert werden: *Faustlos* ist ein Programm zur Förderung sozialer Kompetenzen und zur Reduktion von Gewalt und Aggression, dass sich in der Bundesrepublik Deutschland seit einigen Jahren sehr großer Beliebtheit erfreut. Es basiert auf dem Sozialkompetenzprogramm *Second Step*, welches in den 1980er Jahren von der US-amerikanischen gemeinnützigen Organisation *Committee for Children* entwickelt worden war. Es ist in den USA weit verbreitet (Grossman et al., 1997; Grossman et al., 1987). Nach Einschätzung der deutschen Vertreiber wird auch *Faustlos* inzwischen „in tausenden von Kindergärten und Schulen mit grossem Erfolg eingesetzt". Hierzu gehört beispielsweise, dass die LBS Bayern im Jahr 2004 *Faustlos*- Patenschaften in über 1000 Schulen und Kindergärten Bayerns übernommen hat und diese Patenschaften in den Jahren 2005 und 2006 erneuert hat.

Die Internetseite von *Faustlos* gibt sich bezüglich der erzielten Wirkungen nicht bescheiden. Auf der Frontseite des Internetauftritts wird etwa damit geworben, dass *Faustlos* als *Best-Practice* Projekt in Deutschland gilt. An prominenter Stelle steht dort: *Aufgrund seiner Effektivität und "überregional beispielhaften Qualität" wurde Faustlos im Rahmen der vom Bundesministerium für Bildung und Forschung in Auftrag gegebenen Bestandsaufnahme zu demokratiepolitischen und gewaltpräventiven Potenzialen in Schule und Jugendhilfe als "Best-Practice-Projekt" ausgezeichnet* (siehe www.faustlos.de). Eine Internet-Recherche zeigt, dass dieser Text vielfach übernommen und als Beleg für die Wirksamkeit von Faustlos interpretiert wurde.

Es lohnt sich daher, dieser Auszeichnung genauer nachzugehen. Sie erfolgte im Rahmen einer Expertise und Ausstellung zum Thema *Demokratie lernen in Schule und Gemeinde – demokratiepolitische und gewaltpräventive Potenziale in Schule und Jugendhilfe* (Beutel et al. , 2001). Liest man diesen Bericht, so findet man, dass die Auswahl von insgesamt 35 „best practice" Projekten nicht auf einem Wirkungsnachweis basierte, sondern sehr pragmatisch aufgrund eines *explorierenden Vorgehens* erfolgte (Beutel et al., 2001: 5). Der Bericht nennt Lernqualität, institutionelle Qualität, Aktualität und Originalität als Kriterien.

Allerdings werden diese Kriterien nicht genauer definiert und empirisch überprüfte Effektivität wird nirgends als Auswahlkriterium erwähnt. In der Kurzbeschreibung von *Faustlos*, die man im Anhang des erwähnten Berichtes findet (Beutel et al., 2001: 65-66) wird zudem ausdrücklich darauf hingewiesen, dass zur Wirksamkeit keine Aussagen gemacht werden können, weil entsprechende Auswertungen zum Zeitpunkt der Auszeichnung noch gar nicht vorlagen. Mit anderen Worten: Die Behauptung, die Auszeichnung als „Best-Practice Projekt" sei aufgrund der empirisch nachgewiesenen Wirkungen erfolgt, entspricht nicht der tatsächlichen Sachlage.

Die Programmbetreiber haben inzwischen in zwei deutschen Studien die Wirkungen von *Faustlos* untersucht und die Ergebnisse hierzu auch publiziert (Schick & Cierpka, 2003, 2005). Das Problem bilden hierbei weniger die wissenschaftlichen Publikationen und die dort berichteten Ergebnisse. Problematisch ist die Art und Weise, wie die Ergebnisse öffentlich weiter verbreitet werden. Beispielsweise ist im Faltblatt zu *Faustlos* folgendes zu lesen: „Mit Faustlos liegt ein deutschsprachiges Curriculum vor, das die zentralen gewaltpräventiven Kompetenzen Empathie, Impulskontrolle und den Umgang mit heftigen Gefühlen bei Kindern und Jugendlichen gezielt fördert. Die Effektivität des Programms wurde durch zahlreiche Studien belegt" (Heidelberger Präventionszentrum, 2005). Laien, Politiker und Praktiker müssen aufgrund eines solchen Textes den Eindruck haben, es handle sich bei *Faustlos* um ein Programm mit einem wissenschaftlich abgestützten Nachweis von gewaltpräventiven Wirkungen. Tatsächlich zeigen die Publikationen von Schick und Cierpka (Schick & Cierpka, 2003, 2005) aber etwas völlig Anderes.

Die erste Studie basiert auf einer kontrollierten Experimentalstudie, die zwischen 2001 und 2003 in 44 Grundschulklassen im Raum Heidelberg/Mannheim durchgeführt wurde.[2] Aus der Perspektive der Kinder wurden in dieser Studie sieben Zielgrößen überprüft. Als *wirkungslos* erwies sich *Faustlos* in Bezug auf die Zielgrößen Empathie, Akzeptanz bei anderen Kindern, Selbstvertrauen, Selbstwertgefühl, Angst vor Verletzungen, Angst vor schlimmen Dingen, sowie aggressivem Verhalten. Es konnte ein positiver Effekt auf Angst vor Kontrollverlust gefunden werden.

[2] Die Ergebnisse sind in den Tabellen 2, 3 und 4 von Schick und Cierpka (2003) dargestellt.

Die Angaben der Eltern wurden verwendet, um zwölf Zielgrößen zu prüfen. *Keine präventiven Wirkungen* ergaben sich für folgende Masse: sozialer Rückzug, körperliche Beschwerden, soziale Probleme, schizoid/zwanghaftes Verhalten, Aufmerksamkeitsstörungen, delinquentes Verhalten, aggressives Verhalten, Selbstkontrolle, Selbstbehauptung, Perspektivenübernahme und Kooperation/soziale Regeln. Einzig auf Angst/Depressivität wurde ein knapp signifikanter positiver Effekt gefunden.

Schließlich wurde die Entwicklung der Kinder nach Einschätzung der Lehrpersonen untersucht. Hierbei wurden sechs Zielgrößen analysiert, nämlich: Das Ausmaß von Bandenbildung; die Bereitschaft, anderen zu helfen; Aggression gegen Klassenmitglieder; Diskriminierung gegen Klassenmitglieder; Zusammenhalt zwischen Klassenmitgliedern; Rivalität zwischen Klassenmitgliedern. Für *keine* dieser Zielgrößen berichten die Autoren über eine statistisch abgesicherte Wirkung von *Faustlos*.

Zusammengefasst bedeutet dies, dass für 23 der 25 überprüften Zielgrößen *kein statistisch abgesicherter Effekt* gefunden werden konnte. Hierzu gehören alle Zielgrößen, die sich entweder auf Gewalt selbst oder auf Vorläufer von Gewalt beziehen. Die beiden schwach signifikanten Effekte hingegen wurden in Bereichen beobachtet, die angesichts der Ziele von *Faustlos* marginal sind. Hinzu kommt, dass bei 25 überprüften Effekten rein zufällig ein bis zwei signifikante Wirkungen zu erwarten sind.

In einer zweiten Wirkungsstudie wurde die Version von *Faustlos* an Kindergärten überprüft (Schick & Cierpka, 2004, 2006). An der Studie nahmen 124 Kinder in 14 Kindergärten teil, die nach einem nicht näher erläuterten Verfahren den Experimentalbedingungen zugewiesen wurden. Die Ergebnisse dieser zweiten Studie fallen für die Daten, die direkt bei den Kindern erhoben wurden, positiv aus:[3] Die Autoren können erwünschte Effekte bei unmittelbaren Indikatoren wie der Emotionserkennung, den sozial kompetenten Reaktionen und dem Einsatz von Beruhigungstechniken zeigen, wenn sie aus der Sicht der Kinder selbst beurteilt werden (Schick & Cierpka, 2004: 19). Aggressives Verhalten wurde allerdings aus der Sicht der Kinder nicht gemessen.

Die Autoren finden jedoch keine Wirkungen von *Faustlos*, wenn dieselben Dimensionen aus der Sicht der Eltern beobachtet werden. Die Eltern selbst nehmen keine Verbesserung der emotional-sozialen Kompetenzen des Kindes wahr. Es gibt aus der Sicht der Eltern auch keine Effekte von Faustlos auf Aggressivität oder Ängstlichkeit (Schick & Cierpka, 2004: 20f).

Ebenso wenig hat *Faustlos* aus der Perspektive der Erzieherinnen messbare Effekte auf die Kompetenzen und das Verhalten der Kinder. Dies gilt für alle neun geprüften Verhaltensdimensionen einschließlich Aggressivität (Schick & Cierpka, 2004: 21). In den Verhaltensbeobachtungen zeigt sich kein Effekt von Faustlos auf verbales Verhalten, nonverbale Kompetenz, emotionale Kompetenz und körperliche Aggression. Hingegen finden die Autoren einen kleinen positiven Effekt auf verbale Aggression (Schick & Cierpka, 2004: 22).[4]

Dass bei der Messung von Interventionseffekten durch mehrere Informanden widersprüchliche Ergebnisse gefunden werden, ist leider in der Präventionsforschung eher die Regel als die Ausnahme. Auch erweist es sich in allen Studien zu universeller Gewaltprävention als überaus schwierig, Effekte auf der Ebene des tatsächlichen Kindsverhaltens nachzuweisen. Die Studien zu Faustlos bilden hier keine Ausnahme.

[3] Es ist an dieser Stelle anzufügen, dass die Aussagekraft der Studie durch mehrere methodisch problematische Merkmale eingeschränkt ist. Hiervon seien zwei namentlich erwähnt. Erstens nehmen die Autoren bei der Prüfung der Äquivalenz von Kontroll- und Experimentalgruppe aus nicht näher erläuterten Gründen eine methodisch fragwürdige Korrektur der Signifikanzschwellen vor (Schick und Cierpka, 2004: 14-16). Lässt man diese Korrektur unberücksichtig, dann zeigen die Daten, dass die Kinder in der Experimentalgruppe (im Vergleich zur Kontrollgruppe) vor der Intervention statistisch signifikant jünger waren, dass sie tiefere soziale Kompetenzen hatten und höhere Werte von Problemverhalten nach Eltern- und Lehrereinschätzung aufwiesen. Es kann daher nicht ausgeschlossen werden, dass die Befunde durch Regressionseffekte verzerrt sind. Zweitens wurde bei der Analyse der Effekte die Klumpenrandomisierung nicht berücksichtigt. Eine Klumpenrandomisierung verlangt eine Anpassung der inferenzstatistischen Absicherung. In der vorliegenden Studie dürften daher alle Signifikanzwerte der Interventionseffekte systematisch überschätzt sein.

[4] Allerdings ist darauf hinzuweisen, dass diese Variable vermutlich eine starke Schiefverteilung aufweist, was sich in einer doppelt so hohen Standardabweichung im Vergleich zum Mittelwert ausdrückt (M = 0.015; SD = 0.032 bei einer Skala von 0-1). Es dürfte daher fraglich sein, ob eine konventionelle Varianzanalyse angemessen ist.

Allerdings: Wenn in einem Text der Praktikerzeitschrift „Schulverwaltung" zu lesen ist, das Gewaltpräventionsprogramm Faustlos habe „gerade für ein Präventionsprogramm bemerkenswert große Effekte" (Schick, 2004) erzielt, dann kann der Eindruck nicht völlig von der Hand gewiesen werden, hier werde ein insgesamt doch recht zwiespältiges wissenschaftliches Ergebnis über Gebühr strapaziert und der Präventionspraxis in Bild vermittelt, das durch die Forschung nicht gedeckt ist.

Folgerungen und Empfehlungen

Die wissenschaftspraktische Bewegung der evidenzbasierten Prävention ist in den letzten 20 Jahren mit dem Anspruch angetreten, für Praxis und Politik zuverlässiges Wissen darüber aufbereiten zu können, welche Maßnahmen wirken, welche nicht wirken, und welche schädlich sind. Es ist erfreulich, dass diese Bewegung in den letzten 10 Jahren auch im deutschsprachigen Raum Fuß gefasst hat und neue Anstöße für die Präventionspolitik ausgelöst hat. Es ist ebenfalls zu begrüßen, dass sich Wissenschaftler zunehmend für die Entwicklung von Präventions- und Interventionsprogrammen interessieren und sich der Aufgabe einer Wirkungsevaluation mit Hilfe experimenteller Designs stellen. Angesichts der manchmal sehr hohen Erwartungen der Öffentlichkeit ist allerdings zu beachten, dass dieses Ziel nur dann dauerhaft erreicht werden kann, wenn die Wissenschaft unvoreingenommen und mit der nötigen Vorsicht die möglichst sorgfältig ermittelten Ergebnisse kommuniziert und hierbei keine unrealistischen Erwartungen weckt, welche auf Dauer nicht eingelöst werden können.

Die im vorangehenden Abschnitt gezeigten Einzelbeispiele dokumentieren, dass die Überscheidung von Forschungstätigkeit, Vertrieb von Präventionsprogrammen sowie politischen Beratungsfunktionen nicht unproblematisch ist. Es kommt dabei notwendigerweise zu Interessenkonflikten, die etwa der Situation ähnlich sind, wo ein Jugendlicher Richter über sein eigenes Verhalten sein müsste. Abschließend sollen daher einige Möglichkeiten vorgestellt werden, welche diesen Interessenkonflikt entschärfen können.

Verbindliche Richtlinien für Durchführung und Publikation von Wirkungsevaluationen

Das Programmentwickler ihre eigenen Programme wissenschaftlich evaluieren, ist grundsätzlich zu begrüßen. In dem Ausmaß, in dem in Zukunft mehr standardisierte und in Zusammenarbeit zwischen Forschung und Wissenschaft erarbeitete Präventionsprogramme auf den Markt kommen.

Das Problem der verzerrenden, ungenauen und möglicherweise durch Interessenkonflikte beeinflussten Berichterstattung über Evaluationsstudien wurde in der Medizin deutlich früher als in der Sozialwissenschaft erkannt. Daher wurden in diesem Forschungsbereich schon seit längerem Richtlinien über die Durchführung und Publikationen von Experimentalstudien verfasst, die von wissenschaftlichen Zeitschriften und Fachverbänden als verbindlich betrachtet werden. Was die rein wissenschaftlichen Anforderungen anbelangt, könnten sich Forschende beispielsweise problemlos an den CONSORT (Consolidated Standards for Reporting Trials) orientieren (Altman, 1996). Es enthält alle notwendigen Anweisungen, welche für eine wissenschaftliche Beurteilung notwendig sind. Außerdem existieren spezifisch für die kriminologische Präventionsforschung einschlägige Publikationen, welche Forschenden detaillierte Qualitätsstandards an die Hand geben, welche verschiedene Aspekte der Validität umfassen (Farrington, 2003; Lösel & Koferl, 1989). Besonders hilfreich in dieser Hinsicht ist die Liste von Qualitätskriterien für Evaluationsprojekte, welche Farrington (2003) zusammengestellt hat. Es wäre sinnvoll, wenn auch im deutschsprachigen Raum solche Qualitätsstandards in der praxisorientierten Forschung eine möglichst große Verbreitung finden würden. Fachvereinigungen, Institutionen der Forschungsförderung und öffentliche Auftraggeber könnten beispielsweise Checklisten der Informationen erstellen, die in Forschungsberichten zwingend enthalten sein müssen.

In den USA, wo standardisierte und kommerziell vertriebene Präventionsprogramme schon seit längerem existieren und ein entsprechend vielfältiges Angebot besteht, wurde schon vor gut 10 Jahren die Notwendigkeit erkannt, Praktikern Leitfäden zur Beurteilung von Präventionsprogrammen an die Hand zu geben. Dabei wurde auch erkannt, dass solche Beurteilungen in einem transparenten Reviewverfahren durch unabhängige wissenschaftliche Gremien mit einer hohen Vertrauenswürdigkeit vorgenommen werden sollten.

Modellcharakter in dieser Hinsicht haben die Blueprints of Violence Prevention des Center for the Study and Prevention of Violence an der Universität von Colorado (Elliott & Mihalic, 2004). Diese 1996 begonnene Initiative evaluiert einzelne Präventionsprogramme mit besonderem Blick auf die Frage, ob Programme den Nachweis der Wirksamkeit erbracht haben und sie lokalen Entscheidungsträgern zur Umsetzung empfohlen werden können. Bisher wurden 600 Programme in einem rigiden Verfahren geprüft, hiervon wurden bisher 11 als Modellprogramme eingestuft, weitere 28 gelten gegenwärtig als viel versprechend. Die Programme werden durch ein Gremium von sieben hervorragend qualifizierten Spezialisten der Evaluationsforschung nach standardisierten Kriterien evaluiert. Die wichtigsten dieser Kriterien sind nachweislich positive Effekte in einem rigiden Forschungsdesign, Nachhaltigkeit der Effekte über das Ende der Intervention hinaus sowie Replikation der positiven Effekte in mindestens zwei weiteren, unabhängig durchgeführten Studien.

Angesichts der zunehmenden Popularität von evidenzbasierter Prävention – und der steigenden Zahl von Programmanbietern – wäre es sinnvoll, im deutschen Sprachraum ähnliche Mechanismen der Qualitätskontrolle und praxisnahen Aufbereitung des Wissensstandes zu schaffen. Gegenwärtig gibt es aber weder in der Schweiz noch in Deutschland solche unabhängigen Reviews, welche im Sinne einer Qualitätskontrolle wirken würden.

Unabhängige Evaluationen fördern

Es gibt viele Gründe, warum Evaluationen mit einer starken aktiven Beteiligung der Programmentwickler oder –vertreiber auch in Zukunft die Regel bleiben werden. Hierzu gehört etwa das Interesse an einer wissenschaftlichen Validierung. Es wird in dem Maße weiter zunehmen, als die öffentlichen Nachfrager nach Prävention einen wissenschaftlichen Leistungsausweis zur Bedingung für die Übernahme von Programmen machen. Hinzu kommt, dass der externe Finanzierungsbedarf bei Eigenevaluationen in der Regel sehr viel geringer ist als bei Evaluationen durch Dritte, weil die Programmentwickler oft einen beträchtlichen Teil der Kosten selbst übernehmen.

Dennoch sollten Wissenschaftsräte und Stiftungen gezielt auch unabhängige Evaluationen fördern. Es gibt drei wesentliche Gründe, weshalb unabhängige Evaluationen ein wichtiges Korrektiv zu den Befunden von Selbstevaluationen sind. Erstens ist anzunehmen, dass Programmentwickler besonders sorgfältig auf eine erstklassige Umsetzung, hohe Motivation und überdurchschnittliche Betreuung des Projektes achten. Das ist legitim. Allerdings sind solche Bedingungen in der Alltagspraxis kaum je gegeben. Daher ist es notwendig, Programme auch so zu evaluieren, wie sie auf dem Markt angeboten und im Alltag tatsächlich realisiert werden. Zweitens besteht vermutlich generell in der Präventionsforschung die Tendenz, erwünschte Ergebnisse eher zu publizieren als Nullbefunde oder gar unerwünschte Effekte. Diese auch als Publikationsbias oder „Schubladenproblem" bekannt Phänomen (Rosenthal, 1979) dürfte besonders dann ein Problem sein, wenn für eine Forschergruppe materielle oder immaterielle Interessen mit dem Evaluationsausgang verbunden sind. Drittens dürften Forschende, welche ein Eigeninteresse am geprüften Programm haben, besonders in Versuchung sein, die publizierten Ergebnisse in die gewünschte Richtung zu schönen indem unerwünschte Teilbefunde unterschlagen oder die Ergebnisse durch problematische Manipulationen zurechtgebogen werden.

Bessere Ausbildung von Nutzern in Praxis und Verwaltung

Häufig verfügen die Nutzer von sozialwissenschaftlichen Wirkungsevaluationen nur über sehr begrenztes Wissen, um an eine Studie die richtigen Fragen zu stellen und die Ergebnisse kritisch beurteilen zu können. Es scheint wichtig, dass die verantwortlichen Nutzer von Wirkungsstudien in Politik und Verwaltung über das Fachwissen verfügen, um den wissenschaftlichen Output beurteilen zu können und die richtigen Fragen zu stellen. Es könnte daher sinnvoll sein, ein entsprechendes passives Fachwissen durch Weiterbildungsveranstaltungen und Kurse zu fördern. Fachpersonen in der Praxis sollten verstehen, welche Kriterien die Qualität einer Wirkungsevaluation beeinflussen; welche Forschungsdesigns am besten auf eine Forschungsfrage angemessen sind; welche Informationen sie von einem wissenschaftlichen Bericht erwarten dürfen; und woher sie allenfalls Hilfe bei der Beurteilung der berichteten Effekte einholen können.

Literatur

Altman, D. G. (1996). Better Reporting of Randomised Controlled Trials: the CONSORT Statement. *BMJ, 313*(7057), 570-571.

Asshauer, M., & Hanewinkel, R. (2000). Lebenskompetenztraining für Erst- und Zweitklässler: Ergebnisse einer Interventionsstudie. *Zeitschrift für Klinische Kinderpsychologie, 9*, 251-263.

Beutel, W., Schnurre, S., Senge, K., Thöne, A., & Fauser, P. (2001). *Demokratie lernen in Schule und Gemeinde - Demokratiepolitische und gewaltpräventive Potenziel in Schule und Jugendhilfe*. Berlin: Bundesministerium für Bildung und Forschung.

Borman, G. D., Hewes, G. M., Overman, L. M., & Brown, S. (2003). Comprehensive School Reform and Achievement: A Meta-Analysis. *Review of Educational Research, 73*(2), 125-230.

Campbell, D., & Stanley, J. C. (1966). *Experimental and Quasi-experimental Designs for Research*. Chicago: Rand McNally.

Cochrane, A. (1972). *Effectiveness and Efficiency: Random Reflections on Health Services*. London: Nuffield Provincial Hostpitals Trust.

Cook, T. D., & Campbell, D. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand McNally.

Eisner, M., & Ribeaud, D. (2008). Kritischer Kommentar zu Nina Heinrichs, Kurt Halweg, Heike Bertram, Annett Kuschel, Sebastian Naumann, Sylvia Harstick: Die langfristige Wirksamkeit eines Elterntrainings zur universellen Prävention kindlicher Verhaltensstörungen: Ergebnisse aus der Sicht der Mütter und der Väter. Pädagogisches Institut der Universität Zürich.

Eisner, M., Ribeaud, D., Jünger, R., & Meidert, U. (2007). *Frühprävention von Gewalt und Aggression; Ergebnisse des Zürcher Interventions- und Präventionsprojektes an Schulen*. Zürich: Rüegger.

Ellickson, P. L. (1998). Preventing Adolescent Substance Abuse: Lessons from the Project ALERT Program. In J. Crane (Ed.), *Social Programs That Work* (pp. 201-257). New York: Russell Sage.

Ellickson, P. L., & Bell, R. M. (1990). Drug Prevention in Junior High: A MultisiteLongitudinal Test. *Science, 247*, 1299-1305.

Ellickson, P. L., Bell, R. M., & Harrison, E. R. (1993). Changing Adolescent Propensities to Use Drugs: Results from Project ALERT. *Health Education Quarterly, 20*, 227-242.

Ellickson, P. L., Bell, R. M., & McGuigan, K. (1993). Preventing Adolescent Drug Use: Long-Term Results of a Junior High Program. *American Journal of Public Health, 93*(11), 1830-1836.

Elliott, D., & Mihalic, S. (2004). Issues in Disseminating and Replicating Effective

Prevention Programs. *Prevention Science, 5*(1), 47-53.

Farrington, D. P. (2003). Methodological Quality Standards for Evaluation Research. *Annals of the American Academy of Political and Social Sciences, 587*, 49-68.

Gandhi, A. G., Murphy-Graham, E., Petrosino, A., Chrismer, S. S., & Weiss, C. H. (2007). The Devil is in the Details: Examining the Evidence for "Proven" School-Based Drug Abuse Preventiopn Programs. *Evaluation Review, 31*(1), 43-74.

Grossman, D. C., Neckerman, H. J., Koepsell, T. D., Asher, K. N., Beland, K., Frey, K., et al. (1997). Effectiveness of a Violence Prevention Curriculum Among Children in Elementary School: A Randomized Controlled Trial. *Journal of the American Medical Association, 277*(20), 1605-1611.

Grossman, D. C., Neckerman, H. J., Koepsell, T. D., Liu, P.-Y., Adher, K. N., Beland, K., et al. (1987). Effectiveness of a Violence Prevention Curriculum among Children in Elementary School - A Randomized Controlled Trial. *Journal of the American Medical Association, 277*(20), 1605-1611.

Heidelberger Präventionszentrum. (2005). *Faustlos - Gewaltprävention durch Förderungsozial-emotionaler Kompetenzen*. Heidelberg: Heidelberger Präventionszentrum.

Heinrichs, N., Hahlweg, K., Bertram, H., Kuschel, A., Naumann, S., & Harstick, S. (2006). Die langfristige Wirksamkeit eines Elterntrainings zur universellen Prävention kindlicher Verhaltensstörungen: Ergebnisse aus Sicht der Mütter und Väter. *Zeitschrift für klinische Psychologie und Psychotherapie, 35*(2).

Henggeler, S. W., Cunningham, P. B., Pickrel, S. G., Schoenwald, S. K., & Brondino, M. J. (1996). Multisystemic Therapy: an Effective Violence Prevention Approach for Serious Juvenile Offenders. *Journal of Adolescence, 19*(1), 47-61.

Henggeler, S. W., Melton, G. B., & Smith, L. A. (1992). Family Preservation using Multisystemic Threapy: An Effective Alternative to Incarceration. *Journal of Consulting and Clinical Psychology, 60*(6), 953-961. Littell, J. (2005). Lessons from a Systematic Review of Effects of Multisystemic Therapy. *Children and Youth Services Review, 27*(4), 445-463.

Littell, J., Popa, M., & Forsythe, B. (2005). Multisystemic Therapy for Social, Emotional, and Behavioral Problems in Youth Aged 10-17 (report for the Campbell Collaboration) [Electronic Version]. Retrieved 14 January 2008 from http://www.sfi.dk/graphics/Campbell/Dokumenter/MST_Review/MULTISYSTEMIC%20TERAPHY%20-%20REVIEW.pdf.

Lösel, F., & Beelmann, A. (2003). Effects of Child Skills Training in Preventing Antisocial Behavior: A Systematic Review of Randomized Evaluations. *The ANNALS of the American Academy of Political and Social Science, 587*(1), 84-109.

Lösel, F., Beelmann, A., Stemmler, M., & Jaursch, S. (2006). Probleme des Sozialverhaltens im Vorschulalter: Evaluation des Eltern- und Kindertrainings EFFEKT. *Zeitschrift für klinische Psychologie und Psychotherapie, 35*(2), 127-139.

Lösel, F., & Koferl, P. (1989). Evaluation Research on Correctional Treatment in West Germany: A Meta-Analysis. In H. Wegener, F. Lösel & J. Haisch (Eds.), *Criminal Behavior and the Justice System: Psychological Perspectives*. New York: Springer.

Nachtigall, C., & Suhl, U. (2002). Der Regressionseffekt - Mythos und Wirklichkeit. *m e t h e v a l r e p o r t (http://www.metheval.unijenade/materialien/reports/report_2002_02.pdf, zuletzt aufgerufen am 11.1.2008), 4* (2).

Ogden, T., & Hagen, K. A. (2006). Multisystemic Treatment of Serious Behaviour Problems in Youth: Sustainability of Effectiveness Two Years after Intake. *Child and Adolescent Mental Health, 11*(3), 142-149.

Olsson, T., & Sundell, K. (2008). The Transportability of MST to Sweden: Short-term Results from a Randomized Trial of Conduct Disordered Youth (Manuskript, zur Publikation eingereicht).

Petrosino, A., & Soydan, H. (2005). The Impact of Program Developers as Evaluators on Criminal Recidivism: Results from Meta-analyses of Experimental and Quasiexperimental Research. *Journal of Experimental Criminology, 1*(4), 435-450.

Rosenthal, R. (1979). The File Drawer Problem and Tolerance for Null Results.*Psychological Bulletin, 86*(3), 638-641.

Rössner, D., Bannenberg, B., & Landeshauptstadt Düsseldorf. (2002). *Düsseldorfer Gutachten: Empirisch gesicherte Erkenntnisse über kriminalpräventive Wirkungen*. Düsseldorf: Landeshauptstadt Düsseldorf.

Sanders, M. R. (1999). Triple P-Positive Parenting Program: Towards an Empirically Validated Multilevel Parenting and Family Support Strategy for the Prevention of Behaviour and Emotional Problems in Children. *Clinical Child and Family Psychology Review, 2*(2), 71–89.

Sanders, M. R., Lynch, M. E., & Markie-Dadds, C. (1994). *Every Parent's Workbook: A Practical Guide to Positive Parenting*. Brisbane: Australian Academic Press.

Schick, A. (2004). Inhalte, Implementation und Effektivität eines Gewaltpräventions-Curriculums. *Schulverwaltung Spezial*(3), 22-24.

Schick, A., & Cierpka, M. (2003). Faustlos - Evaluation eines Curriculums zur Förderung sozial/emotionaler Kompetenzen und zur Gewaltpraevention in der Grundschule. *Kindheit und Entwicklung, 12*(2), 100-110.

Schick, A., & Cierpka, M. (2004). *Evaluation des Faustlos-Curriculums für den Kindergarten (Schriftenreihe der Landesstiftung Baden-Württemberg, Nr 7)*. Retrieved 14 January 2008, from http://www.landesstiftungbw.de/publikationen/files/schr_eval_faustkinder.pdf.

Schick, A., & Cierpka, M. (2005). Faustlos: Evaluation of a Curriculum to Prevent Violence in Elementary Schools. *Applied and Preventive Psychology, 11*(1), 157-165.

Schick, A., & Cierpka, M. (2006). Evaluation des Faustlos-Curriculums für den Kindergarten. *Praxis der Kinderpsychologie und Kinderpsychiatrie, 55*(6), 157-165.

Shadish, W. R., Cook, T. D., & Campbell, D. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.

Sherman, L. W., Farrington, D. P., Welsh, B. C., & MacKenzie, D. L. (Eds.). (2002). *Evidence-Based Crime Prevention*. London: Routledge.

St Pierre, T. L., Osgood, D. W., Mincemoyer, C. C., Kaltreider, D. L., & Kauh, T. J. (2006). Results of an Independent Evaluation of Project ALERT Delivered in Schools by Cooperative Extension. *Prevention Science, 6*(4), 305-317.

Timmons-Mitchell, J., Bender, M. B., Kishna, M. A., & Mitchell, C. C. (2006). An Independent Effectiveness Trial of Multisystemic Therapy With Juvenile Justice Youth. *Journal of Clinical Child & Adolescent Psychology, 35*(2), 227-236.

Problematische Interpretation eines randomisierten Feldversuchs: Eine methodische Kritik von Nina Heinrichs et al. (2006) „Die langfristige Wirksamkeit eines Elterntrainings zur universellen Prävention kindlicher Verhaltensstörungen: Ergebnisse aus der Sicht der Mütter und der Väter" ,in: *Zeitschrift für klinische Psychologie und Psychotherapie,* 35(2), 97-108.

Der Beitrag von Heinrichs et al. (2006) berichtet über eine Studie zur Wirksamkeit des Elterntrainings Triple P als universelles Programm zur Prävention kindlichen Problemverhaltens. Die Autoren kommen in dem Beitrag zum Schluss, dass ihre Studie die Wirksamkeit von Triple P als universellem Elterntraining zeige und entsprechend „an eine breitflächige Dissemination dieses Trainings gedacht werden" solle. Wie wir im Folgenden darlegen, weist die vorgestellte Studie allerdings eine Reihe gravierender methodischer Fehler auf, die eine wissenschaftliche Beurteilung der tatsächlich erreichten Effekte völlig verunmöglichen. In den Schlussfolgerungen diskutieren wir die Problematik der vorliegenden Studie vor dem Hintergrund neuer Erkenntnisse der Evaluationsforschung.

Intention-to-treat

Die Autoren beanspruchen in dem Text mehrfach, die Daten gemäss einem "Intention-to-treat" (ITT) Ansatz auszuwerten (Heinrichs et al., 2006: 89, 91, 92). Dieser Anspruch ist irreführend. In der relevanten wissenschaftlichen Literatur herrscht Einigkeit darüber, was unter einer ITTAnalyse zu verstehen ist (Fisher et al., 1990; Newell, 1992). Um nur eine Quelle zu zitieren: "Intention to treat analysis includes all randomized patients in the groups to which they were randomly assigned, regardless of their adherence with the entry criteria, regardless of the treatment they actually received, and regardless of subsequent withdrawal from treatment or deviation from the protocol" (Fisher et al., 1990). Genau gleich definiert die Cochrane Collaboration auf ihrer Webseite: "The basic intention-to-treat principle is that participants in trials should be analysed in the groups to which they were randomized, regardless of whether they received or adhered to the allocated intervention."

In randomisierten Experimentalstudien wird eine ITT-Analyse aus zwei Gründen als die optimale Auswertungsstrategie angesehen: Erstens bildet jene Teilgruppe der Behandlungsgruppe, die wirklich an einer Behandlung teilnimmt, keine echte Zufallsstichprobe mehr, so dass die Analysen durch Selbstselektionseffekte verzerrt sein können. Zweitens ist für praktische Zwecke wichtig zu wissen, wie gross der Behandlungseffekt insgesamt ist, und nicht nur unter denjenigen, welche sich vollständig der Behandlung unterziehen. Mit anderen Worten wird damit die Gesamtwirksamkeit unter Berücksichtigung der „Compliance" der Behandlung gemessen.

Was in dem Beitrag von Heinrichs und Kollegen getan wird, hat allerdings hiermit nichts zu tun. Entscheidend ist die kurze Passage auf Seiten 86/87: „Ein Ziel der vorliegenden Studie war es, differenzielle Unterschiede in der Effektivität des Triple P-Gruppentrainings zwischen Müttern und Vätern zu untersuchen. Daher wurden nur die Zwei-Eltern-Familien ($N = 219$, davon $N = 129$ Triple P) in die Auswertung aufgenommen. Es zeigten sich keine signifikanten Unterschiede sowohl in den sozioökonomischen als auch Kriterien-Variablen zwischen den Familien, die zur Kontrollbedingung randomisiert wurden ($N = 62$) und den Triple P-Ablehnern ($N = 28$), so dass beide Gruppen für die folgende Auswertung zusammengefasst wurden ($N = 90$)." Im Klartext heisst das: Für die statistische Analyse wurden *nachträglich* 28 Eltern aus der ursprünglichen Behandlungsbedingung in die Kontrollgruppe umgeteilt. Bei ursprünglich 129 Eltern in der Experimentalgruppe und 62 Eltern in der Kontrollgruppe ist diese Massnahme alles andere als trivial. Methodisch ist das Vorgehen in keiner Weise zu rechtfertigen. Und mit Sicherheit darf nach einer solchen Umteilung nicht mehr von einer ITT-Analyse gesprochen werden. Über die Gründe für die Manipulation macht der Text keine Angaben, so dass man darüber nur spekulieren kann. Denkbar ist, dass hiermit statistische Power gewonnen werden sollte, liegen doch die berichteten Effekte auf kindliches Problemverhalten alle nahe an der Grenze zur statistischen Nicht-Signifikanz.

Kontrollgruppe und Experimentalgruppe

Der zentrale Grund für randomisierte Kontrollgruppenstudien besteht darin, dass Experimentalgruppe und Kontrollgruppe hinsichtlich aller Hintergrundsvariablen sowie hinsichtlich der Kriteriumsvariablen zum Zeitpunkt der Nullmessung als gleichwertig angesehen werden können. Entscheidend ist, dass alle Variablen, die möglicherweise einen Effekt auf die Zielgrösse haben können, als gleich verteilt angenommen werden können. Entsprechend wichtig ist eine transparente Diskussion der Frage, ob durch die Randomisierung Äquivalenz erreicht werden konnte.

Die Autoren suggerieren, dass sie den Nachweis für die Äquivalenz der Kontroll- und der Experimentalgruppe in Tabelle 1 erbringen. Insbesondere sagen sie, „bezüglich der sozioökonomischen Merkmale zeigten sich „keine signifikanten Unterschiede zwischen der Kontroll- und der der Experimentalgruppe".

Diese Behauptung ist – erstens – irreführend. Wie man nämlich den Angaben zum *N* entnehmen kann, beziehen sich die in Tabelle 1 gezeigten Daten nicht auf die ursprünglichen Experimental bzw. Kontrollgruppen, sondern auf die Gruppen *nach* der oben erwähnten Umteilung. Äquivalenz nach einer nachträglichen Neuzuteilung ist aber für die Beurteilung einer gelungenen Randomisierung absolut belanglos.

Weit gravierender ist allerdings – zweitens –, dass in Tabelle 1 jeglicher Hinweis auf die Frage der Äquivalenz hinsichtlich der *Kriteriumsvariablen* fehlt. Man kann die Antwort hierauf der Tabelle 2 entnehmen. Man sieht dort, dass *alle* Indikatoren zum Problemverhalten des Kindes in der Experimentalgruppe massiv höher sind als in der Kontrollgruppe, also *keine Äquivalenz vorliegt*. Hinsichtlich des Gesamtwertes des CBCL etwa liegt der Mittelwert in der Experimentalgruppe zum Prä-Zeitpunkt fast sechs Punkte höher als in der Kontrollgruppe (33.14 versus 26.34). Das entspricht fast einer halben Standardabweichung.

Diese Daten zeigen, dass in der vorliegenden Studie die wichtigste Annahme eines randomisierten Kontrollgruppendesigns verletzt ist. Entsprechend dürfen die in Tabelle 3 berichteten Unterschiede zwischen den Gruppen hinsichtlich der Prä-Post-Differenzen *nicht* als Beleg für erfolgreiche Interventionseffekte interpretiert werden. Auf die Gründe dafür, warum trotz des randomisierten Vorgehens derart grosse Unterschiede in den Ausgangswerten beobachtet werden, wird in dem Text mit keinem Wort eingegangen. Es liegt nahe anzunehmen, dass sich v.a. Eltern mit hoher Problemwahrnehmung in die Behandlungsgruppe „selegiert" haben, während es wahrscheinlich scheint, dass sich insbesondere Eltern mit geringer Problemwahrnehmung gegen einen Kursbesuch entschieden und so in die Kontrollgruppe umgeteilt wurden.

Nur am Rande sei hier darauf hingewiesen, dass ausserdem eine unhinterfragte Analyse der Individualdaten noch aus einem zweiten Grund äusserst problematisch ist: Wie man den Schilderungen zum Forschungsdesign entnehmen kann, wurden in dieser Studie nicht Individuen sondern Kindergärten randomisiert. Bekanntlich verlangt eine gruppenweise Randomisierung aber entsprechende Analyseverfahren, mindestens aber eine Prüfung der Varianzanteile, die innerhalb der Cluster gebunden sind (Cook, 2005).

Regressionseffekte

Der Begriff „Regression" wurde von Galton (1886) aufgrund der Beobachtung entwickelt, dass sich Gruppen mit unterschiedlichen Ausgangswerten im Verlauf der Zeit aufeinander zu bewegen. Dass in Experimentalstudien, in denen die Behandlungsgruppen unterschiedliche Ausgangswerte haben, mit einer solchen „Regression to the mean" zu rechnen ist, ist aus der Forschung gut belegt (Bland und Altman, 1994). Auch Lösel et al. (2006) weisen in ihrem Beitrag zum Themenheft, in welchem auch die Arbeit von Heinrichs et al. erschienen ist, ausdrücklich auf dieses Problem hin.

Genau dies scheint nun in der Studie von Heinrichs et al. der Fall zu sein. So kann man Tabelle 2 problemlos entnehmen, dass sich alle Indikatoren von Problemverhalten aufeinander zu bewegen. Für die Gesamtskala des CBCL etwa sinkt die Differenz von sechs Punkte in der Prä-Messung auf rund 2.5 Punkte in der Post-Messung. Dass dieser Datenverlauf ohne weiteres als Regressionseffekt interpretiert werden kann, wird nur am Rande

erwähnt. Im Wesentlichen wird aber in den Folgerungen und im Abstract beansprucht, die Daten würden die Wirksamkeit von Triple P als universelle Intervention dokumentieren.

Angesichts des Fehlens von Äquivalenz bei der Nullmessung würde es sich aufdrängen, mittels einer Kovarianzanalyse mindestens zu prüfen, welche Kovariaten mit den Unterschieden bei der Prä-Messung zusammenhängen und diese in der Beurteilung der Effekte angemessen zu berücksichtigen.

Teilnahmerate und konservative Schlussfolgerungen

Die Autoren geben mehrfach ihrer Meinung Ausdruck, dass die geschätzten Effektstärken als „sehr konservativ" zu beurteilen seien (z.B. S. 92 und 94). Diese Einschätzung ist bereits wegen der oben diskutierten Verletzung des ITT-Designs irreführend. Zweitens muss hervorgehoben werden, dass in der vorliegenden Studie 69 Prozent der angefragten Eltern sich nicht nur nicht an der Intervention beteiligten, sondern ganz generell die *Studienteilnahme verweigert* haben. Natürlich ähneln bei einer derart tiefen Teilnahmerate Rückschlüsse auf die Gesamtpopulation, aus der die Stichprobe gezogen wurde, generell einer Lotterie. Wenn überhaupt, dann müsste man angesichts dieser Situation eher darauf hinweisen, dass man nur hoch motivierte Mittelschichteltern erreicht hat, bei denen möglicherweise die Effekte überschätzt werden.

Weit wichtiger ist aber, dass häufig die Teilnehmenden einer Präventionsmassnahme deren Wirkungen überschätzen. Dass solche Placebo-Effekte bei sozialwissenschaftlichen Experimenten auftreten können, ist gut bekannt und in der Fachliteratur ausführlich diskutiert. Daher wird in der Literatur zur Wirkungsforschung auch regelmässig gefordert, dass Effekte mit einem Multi-Informanten Ansatz geprüft werden müssen. In der vorliegenden Studie wird nur über die Einschätzungen der Väter berichtet. Diese haben sich weit überwiegend nicht an der Intervention beteiligt und können auch keine Verhaltensänderung der Kinder erkennen. Dieses Muster könnte problemlos dadurch erklärt werden, dass die Mütter mehr Zeit in die Intervention investiert haben und daher eher einer Selbsttäuschung unterliegen.

Schlussfolgerungen

Aus unserer Sicht wurden in diesem Beitrag grundlegende methodologische Prinzipien verletzt. Dies ist umso erstaunlicher, als der Unterschied zwischen individueller und gruppenbasierter Randomisierung, das Prinzip der „Intention-to-Treat" Analyse und die Folgen ungleichwertiger Teilgruppen, die Auswirkungen von Regressionseffekten, sowie die Bedeutung der Messung von Zielgrössen durch verschiedene Beobachter heute zum Standardwissen in der Präventionsforschung gehören. Angesichts dieser Probleme stehen die Folgerungen der Autoren leider auf überaus schwachen Füssen. Jedenfalls kann keine Rede davon sein, dass die Studie, wie sie im vorliegenden Beitrag präsentiert wird, die Wirksamkeit von Triple P als universelles Elterntraining belegt.

Wir glauben, dass der vorliegende Beitrag vor dem Hintergrund eines allgemeinen Problems zu betrachten ist. So wurde in den letzten Jahren immer deutlicher, dass Evaluationsstudien durchgehend deutlich höhere Effekte berichten, wenn sie von Forschenden realisiert wurden, welche selbst an Entwicklung und Vertrieb des Präventionsprogramms beteiligt sind, als wenn sie in unabhängigen Studien geprüft wurden. Oft zeigt sich gar, dass unabhängige Studien die Effekte überhaupt nicht (Gandhi et al., 2007) oder nur in sehr beschränktem Mass (St. Pierre et al., 2006) replizieren können. Hierfür kann es verschiedene Gründe geben: Erstens ist anzunehmen, dass Programmentwickler besonders sorgfältig auf eine erstklassige Umsetzung, hohe Motivation und überdurchschnittliche Betreuung des Projektes achten. Das ist legitim. Allerdings sind solche Bedingungen in der Alltagspraxis kaum je gegeben. Daher ist es notwendig, Programme auch so zu evaluieren, wie sie auf dem Markt angeboten und im Alltag tatsächlich realisiert werden. Zweitens besteht vermutlich generell in der Präventionsforschung die Tendenz, erwünschte Ergebnisse eher zu publizieren als Nullbefunde oder gar unerwünschte Effekte. Diese auch als Publikationsverzerrung oder „Schubladenproblem" bekannt Phänomen (Rosenthal, 1979) dürfte besonders dann ein Problem sein, wenn für eine Forschergruppe materielle oder immaterielle Interessen mit dem Evaluationsausgang verbunden sind. Drittens dürften Forschende, welche ein Eigeninteresse am geprüften Programm haben, besonders in Versuchung sein, die publizierten Ergebnisse in die

gewünschte Richtung zu schönen, indem unerwünschte Teilbefunde unterschlagen oder die Ergebnisse durch problematische Manipulationen zurechtgebogen werden.

Entsprechend wichtig scheint es uns, Präventionsprogramme unabhängigen Evaluationen zu unterziehen.

Literatur

Bland, J. M. & D. G. Altman (1994). "Statistic Notes: Regression towards the mean." *British Medical Journal*, 308(6942): 1499.

Cook, T. D. (2005). "Emergent principles for the design, implementation and analysis of clusterbased experiments in social science." *Annals of the American Academy of Political and Social Sciences*, 599: 176-198.

Fisher, L. D., D. O. Dixon et al. (1990). Intention to treat in clinical trials. In: K. E. Peace (ed.): *Statistical issues in drug research and development.* New York: Marcel Dekker.

Galton, F. (1886). "Regression Towards Mediocrity in Hereditary Stature." *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15: 246-263.

Gandhi, A. G., E. Murphy-Graham et al. (2007). "The Devil is in the Details: Examining the Evidence for "Proven" School-Based Drug Abuse Preventiopn Programs." *Evaluation Review*, 31(1): 43-74.

Heinrichs, N., K. Hahlweg et al. (2006). "Die langfristige Wirksamkeit eines Elterntrainings zur universellen Prävention kindlicher Verhaltensstörungen: Ergebnisse aus Sicht der Mütter und Väter." *Zeitschrift für klinische Psychologie und Psychotherapie*, 35(2): 97-108.

Lösel, F., A. Beelmann et al. (2006). " Prävention von Problemen des Sozialverhaltens im Vorschulalter: Evaluation des Eltern- und Kindertrainings EFFEKT." *Zeitschrift für klinische Psychologie und Psychotherapie*, 35(2): 127-139.

Newell, D. J. (1992). "Intention-to-Treat Analysis: Implications for Quantitative and Qualitative Research." *International Journal of Epidemiology*, 21(5): 837-841.

Rosenthal, R. (1979). "The File Drawer Problem and Tolerance for Null Results." *Psychological Bulletin*, 86(3): 638-641.

St. Pierre, T. L. D. W. Osgood et al. (2006). "Results of an Independent Evaluation of Project ALERT Delivered in Schools by Cooperative Extension." *Prevention Science*, 6(4): 305-317.

Die „Eltern und Schule stärken Kinder" (ESSKI) Studie – Eine Reanalyse

Zusammenfassung

In diesem Arbeitspapier wird eine Reanalyse der Ergebnisse der *Eltern und Schule stärken Kinder* (ESSKI) Studie vorgenommen. Ein im Januar 2007 publizierter Forschungsbericht kam zum Schluss, dass die im Rahmen der Studie realisierten Interventionen wirksam zu einer Reduktion von Aggression und Hyperaktivität geführt hätten; dass das subjektive Wohlbefinden der Kinder gestärkt worden sei; und dass die Massnahmen die Wahrscheinlichkeit von Rauchen reduziert hätten. In diesem Arbeitspapier argumentiere ich, dass diese Folgerungen aus den vorliegenden Daten nicht abgeleitet werden können. Vielmehr zeigen die Reanalysen keinen systematisch positiven Effekt der Interventionen auf zentrale Zielgrössen. Von 27 geprüften Effekten fielen 15 zu Gunsten der Kontrollgruppe und 9 zu Gunsten einer der Interventionsgruppen aus. Nur zwei Effekte waren statistisch signifikant, einer zugunsten der Kontrollgruppe, einer zu Gunsten der Interventionsgruppe. In den Schussfolgerungen leite ich hieraus ab, dass Fachverbände strengere Richtlinien über die Präsentation von sozialwissenschaftlichen Wirkungsstudien in der Öffentlichkeit erstellen und durchsetzen sollten.

Hintergrund

Programme zur Frühprävention von Verhaltensproblemen, Gewalt und Suchtverhalten sind während den vergangenen zehn Jahren in der Schweiz und in Deutschland immer beliebter geworden. Ein besonderes Interesse besteht an Programmen, welche entweder elterliche Erziehungskompetenzen fördern oder welche in den Schulen soziale und kognitive Kompetenzen unterstützen. Zunehmend beanspruchen solche Programme auch, über einen wissenschaftlich fundierten Wirkungsnachweis zu verfügen, idealerweise auf der Grundlage eines randomisierten kontrollierten Experimentaldesigns, dem Goldstandard der Präventionsforschung (Shadish, Cook, & Campbell, 2002).

Allerdings ist es der breiteren Öffentlichkeit, den Medien, der Politik oder der Verwaltung oft kaum möglich, den Realitätsgehalt der behaupteten Wirkungen zu überprüfen. Daher ist es notwendig, dass die Ergebnisse von Experimentalstudien möglichst genau und differenziert dargestellt werden und gegebenenfalls von Dritten kritisch diskutiert werden. Dies ist umso wichtiger, als im ganzen Präventionsbereich Evaluationen in der Regel von den Programmentwicklern oder –vertreibern selbst vorgenommen werden. Bei solchen Selbstevaluationen liegt notwendigerweise ein Konflikt zwischen den legitimen Interessen als Programmentwickler oder Lizenznehmer und den Bestrebungen als neutraler Wissenschaftler vor.

Dies muss nicht heissen, dass wissenschaftliche Selbstevaluationen durch Programmentwickler von vornherein problematisch sind. Es gibt aber inzwischen mehrere Arbeiten, die zeigen, dass unabhängige Evaluationen die von Entwicklern behaupteten Präventionswirkungen nicht replizieren konnten (für Drogenprävention vgl. z.B. St Pierre, Osgood, Mincemoyer, Kaltreider, & Kauh, 2006) oder dass wissenschaftlich kaum vertretbare Entscheidungen der Forschenden – etwa das Nicht-publizieren von unerwünschten Befunden - zu einem übermässig positiven Bild der erzielten Wirkungen führen können (für Multisystemische Therapie etwa Littell, 2005).

Ausserdem zeigt die Evaluationsforschung, dass bei Forschungen, welche von den Vertreibern eines Programms vorgenommen werden, die publizierten Effekte im Durchschnitt deutlich höher ausfallen als bei Forschungen, welche von unabhängigen Forschenden realisiert werden (Farrington & Welsh, 2003; Reyno & McGrath, 2006; St Pierre et al., 2006; Wilson, Lipsey, & Derzon, 2003). Die Gründe hierfür werden kontrovers beurteilt. Aber es besteht Einigkeit, dass es bei Eigenevaluationen besonders angebracht ist, die Befunde kritisch zu beleuchten.

Schliesslich haben in der Schweiz und Deutschland Forschungen zu evidenzbasierter Prävention erst vor wenigen Jahren ernsthaft begonnen (Eisner, Ribeaud, & Bittel, 2006). Sie bildet einen potentiell wichtigen und wertvollen

Beitrag zur besseren Allokation von öffentlichen Mitteln im Präventionsbereich. Allerdings gibt es gegenwärtig noch kaum Standards für die wissenschaftliche Berichterstattung und die öffentliche Verbreitung von Forschungsbefunden. Um öffentliche Glaubwürdigkeit für diesen wichtigen Teil praxisbezogener Forschung zu erreichen – und um spätere Enttäuschungen zu vermeiden - ist es sinnvoll, entsprechende Publikationen genau zu betrachten.

Die ESSKI-Studie

In diesem Arbeitspapier unterziehe ich eine der grössten bisher in der Schweiz durchgeführten randomisierten Feldstudien zur Frühprävention von Problemverhalten einer kritischen Reanalyse. Es handelt sich das Projekt *Eltern und Schule stärken Kinder*, ESSKI.

ESSKI ist eine Feldstudie mit einem randomisierten Kontrollgruppendesign zur Wirksamkeit von Massnahmen in Schule und Familie zur „Förderung der Gesundheit von Lernenden, Lehrpersonen und Eltern und zur Prävention von Sucht, Gewalt und Stress" (ESSKI-Projektbeschrieb). Die zentralen realisierten Präventionsmassnahmen sind *Triple P* als Elterntraining zum Selbststudium (Sanders, Markie-Dadds, & Turner, 2003) sowie das Programm *Fit und Stark fürs Leben* als Sozialkompetenzprogramm für die Kinder (Burow, Asshauer, & Hanewinkel, 1998). Die wissenschaftliche Studie wurde durch ein Team von Forschenden der Fachhochschule Nordwestschweiz, der Universität Fribourg, der Pädagogischen Hochschule Zürich und des Schweizerischen Fachstelle für Alkohol- und andere Drogenprobleme realisiert. Bei beiden Massnahmen handelt es sich um Präventionsprogramme, die von Mitgliedern des Forschungsteams selbst in der Schweiz vertreiben werden.

Der Schlussbericht der Studie wurde am 18. Januar 2007 öffentlich vorgestellt.[5] Dem Schlussbericht und der Medienmitteilung kann man folgende Aussagen zur Wirksamkeit der realisierten Massnahmen entnehmen (Eltern und Schule stärken Kinder, 2007; Schönenberger, Lattmann et al., 2006; Schönenberger, Schmid et al., 2006):

„Kinder und Jugendliche, die am Gesundheitsförderungsprojekt ESSKI teilgenommen haben, sind weniger aggressiv und hyperaktiv, fühlen sich gesünder und rauchen weniger" (Medienmitteilung zur ESSKI-Studie vom 18. Januar 2007, erster Satz).

„Es zeigte sich, dass die Schülerinnen und Schüler der Interventionsgruppen weniger streitsüchtig, aufbrausend und hyperaktiv waren, sich seltener unglücklich und niedergeschlagen fühlten und weniger über körperliche Beschwerden klagten." (Medienmitteilung zur ESSKI-Studie vom 18. Januar 2007)

„Die positiven Resultate […] belegen, dass Gesundheitsförderung und Suchtprävention, die schon auf der Primarstufe einsetzen und alle Beteiligten, d.h. auch Eltern und Lehrpersonen einbeziehen, wirksam sind." (Medienmitteilung zur ESSKI-Studie vom 18. Januar 2007)

„Die Auswertungen zum Projekt ESSKI haben ergeben, dass die Stärken der Schülerinnen und Schüler in der Beurteilung der Lehrpersonen und der Eltern zu- und die Schwächen abnehmen." (ESSKIStudie, S. 2)

Diese positive Darstellung der wissenschaftlichen Ergebnisse wird an keiner Stelle relativiert, so dass der Leser den Eindruck erhalten muss, in der Studie seien durchwegs wissenschaftlich abgesicherte positive Ergebnisse erzielt worden. Solche Ergebnisse sind im internationalen Forschungskontext erstaunlich. Uneingeschränkt positive Ergebnisse sind bei universellen Interventionen äussert selten und werden nicht einmal bei sehr intensiven Programmen mit einer hohen Interventionsdosis erreicht. Die Regel ist, dass universelle Programme, wenn sie überhaupt Wirkungen nachweisen können, Teileffekte erzielen und dass diese Teileffekte oft schon Monate nach der Intervention nicht mehr nachweisbar sind (vgl. z.B. Lösel & Beelmann, 2003; Wilson & Lipsey,

[5] Der Text der *Medienmitteilung vom 18. Januar 2007* kann im Internet von der Webseite des Projektes runtergeladen werden. http://www.esski.ch/html/download.htm, Der wissenschaftliche Bericht *Eltern und Schule stärken Kinder: Ein Projekt zur Förderung der Gesundheit bei Lehrpersonen, Kindern und Eltern und zur Prävention von Stress, Aggression und Sucht* (im folgenden: ESSKI-Studie) ist im Internet unter derselben Adresse öffentlich verfügbar.

2006; Wilson et al., 2003). Sollten daher die Befunde der ESSKI Studie wahr sein, so kämen sie einer kleineren wissenschaftlichen Sensation gleich.

Das vorliegende Dokument schildert mein Vorgehen bei der Reanalyse der vorliegenden Daten. Es zeigt, dass eine Reanalyse zu Schlüssen führt, welche von jenen der Projektleitenden in fast allen Punkten abweichen. Es diskutiert ausserdem mögliche Gründe, warum ich bezüglich der inhaltlichen Interpretation zu anderen Schlüssen gelange.

Es mag ungewöhnlich sein, Forschungsergebnisse nicht anhand von Publikationen in Fachzeitschriften sondern auf der Grundlage von Forschungsberichten und Medienmitteilungen zu diskutieren. Allerdings sollte beachtet werden, dass besonders in der praxisorientierten Forschung die Ergebnisse und Folgerungen, welche auf Internetseiten und via Medien verbreitet werden, ein zunehmend wichtiges Instrument der Wissenschaftskommunikation bilden. Entsprechend sollten sie als wissenschaftliche Outputs Ernst genommen und auch kritisch beleuchtet werden.

Vorgehen bei der Reanalyse der Ergebnisse

Der folgenden Reanalyse liegen drei Prinzipien zugrunde: 1. Es werden Zielgrössen untersucht, welche die Autoren der Studie selbst als zentrale Outcomes bezeichnen. Damit wird sicher gestellt, dass die Reanalyse den Zielen des Projektteams gerecht wird. 2. Es werden für die Zielbereiche *alle* Teilaspekte geprüft, über welche in den Berichten Daten vorliegen. Damit wird erreicht, dass es zu keiner Verzerrung zu Gunsten oder zu Ungunsten der Studie kommt. 3. Es werden allgemein anerkannte und robuste Verfahren zur Schätzung von Effektstärken aufgrund publizierter Daten eingesetzt. Damit wird erreicht, dass die nachfolgend dargestellten Ergebnisse auch für Laien verständlich und durch Fachpersonen aufgrund der publizierten Tabellen überprüfbar sind.

Als Grundlage dienen zwei wissenschaftliche Berichte: die schon erwähnte deutschsprachige Studie *Eltern und Schule stärken Kinder: Ein Projekt zur Förderung der Gesundheit bei Lehrpersonen, Kindern und Eltern und zur Prävention von Stress, Aggression und Sucht* sowie der englischsprachige Bericht *Empowerment in family and school (eifas): A randomized trial* (im Folgenden: Empowerment-Studie). Der letztgenannte Bericht ist nicht veröffentlicht und wurde mir durch das Projektteam zur Verfügung gestellt. Er ist für die folgenden Analysen wichtig, weil der Anhang tabellarische Auswertungen zu Teilbereichen enthält, welche im deutschsprachigen Bericht aus nicht näher genannten Gründen weggelassen wurden.

Inhaltliches Vorgehen

Die Autoren der Studie heben drei Zielgrössen hervor, in denen sie nach eigenen Angaben positive Effekte erzielt haben. Dies sind (siehe zitierte Textausschnitte oben): *Weniger Schwächen und mehr Stärken der Schüler; eine bessere Lebenszufriedenheit und weniger körperliche Beschwerden der Schüler*, sowie ein *geringeres Raucherrisiko*. Zur Prüfung dieser Effekte wurden in den Befragungen der ESSKI Studie folgende Instrumente eingesetzt:

Zur Messung von prosozialem Verhalten und Verhaltensstörungen verwendeten die Autoren den *Strengths and Difficulties Questionnaire* von Goodman (2001), einem Befragungsinstrument mit 25 Fragen, das bei 3 bis 16-Jährigen eingesetzt werden kann. Goodman selbst unterscheidet fünf Teilskalen mit je fünf Fragen. Die Autoren der ESSKI Studie hingegen beschränken sich auf zwei Gesamtskalen „Schwächen" und „Stärken". Die Teilskala für „Schwächen" umfasst emotionale Probleme, Verhaltensprobleme und Hyperaktivität. Die Teilskala für „Stärken" misst prosoziales Verhalten. Die Autoren bilden folgende Teilskalen:
- Schwächen der Kinder, Einschätzung der Eltern
- Stärken der Kinder, Einschätzung der Eltern
- Schwächen der Kinder, Einschätzung der Lehrperson
- Stärken der Kinder, Einschätzung der Lehrperson

Zur Erhebung der Lebensqualität der Kinder wurde der Kid-Kindl Questionnaire von Bullinger, von Mackensen und Kirchberger (1994) und Ravens-Sieberer (1998) eingesetzt. Das Originalinstrument umfasst sechs Teildimensionen, nämlich körperliches Wohlbefinden, psychisches Wohlbefinden, Selbstwertgefühl der Kinder, gute Beziehungen zu den Eltern, gute Beziehungen zu anderen Kindern sowie Spass am Unterricht. Jede Teildimension wird mit vier Fragen bewertet. In den mir zugänglichen Berichten werden nur die Ergebnisse von den fünf letztgenannten Teildimensionen berichtet, aber nicht jene zum körperlichen Wohlbefinden. Ich habe keine Anhaltspunkte dafür, ob dieser Aspekt nicht gemessen wurde oder ob die entsprechenden Ergebnisse nicht dargestellt wurden.

Zu Nikotinkonsum und Einstellungen zum Rauchen werden drei Fragen gestellt. Die Fragen sind Einzelitems und es gibt keine Quellenangabe zur Herkunft der Fragen. Die Fragen betreffen den Wunsch, später ein Raucher zu werden; die Überzeugung, dass Rauchen cool sei; sowie die Frage, ob das Kind gegenwärtig rauche.

Ich verwende im Folgenden alle Instrumente wie sie von den Autoren konstruiert wurden. Leider lässt sich für keine der Skalen beurteilen, ob sie eine ausreichende *Skalenreliabilität* aufweist, da die Berichte hierzu keine Informationen enthalten. Ebenso werde ich nicht weiter die *Validität der Messungen* diskutieren. Es sei hier nur angemerkt, dass der *Strengths and Difficulties Questionnaire* nur zwei Fragen zu aggressiven Verhalten hat („hat oft Wutanfälle, ist aufbrausend"; „streitet sich oft mit anderen Kindern oder schikaniert sie") und sich keine der beiden Fragen auf körperliche Gewalt bezieht. Es ist fraglich, ob zwei Fragen ausreichen, um das Konstrukt „Aggression" valide zu messen.

Methodisches Vorgehen

Der Anhang der beiden oben erwähnten Berichte enthält Daten zu den Messungen zum Prä-Zeitpunkt (unmittelbar vor der Intervention), zum Post-Zeitpunkt (unmittelbar nach der Intervention) und zum Follow-Up-Zeitpunkt (fünf Monate nach der Intervention).

Ich beschränke mich im Folgenden auf eine Analyse der Veränderungen zwischen Prä-Zeitpunkt und Follow-Up Zeitpunkt. Der Grund hierfür ist, dass aus praktischer Sicht Wirkungen kaum von Bedeutung sind, wenn sie nicht wenigstens noch nach fünf Monaten beobachtet werden können. Ausserdem kann hierdurch die Fragestellung einfach gehalten werden. Sie lautet: War in einer der Interventionsgruppen die Veränderung statistisch signifikant besser als in der Kontrollgruppe?

Diese Frage lässt sich allerdings nur für die beiden Bereiche *Stärken und Schwächen* und *Lebensqualität* untersuchen. Für die Fragen zum Rauchen hingegen dokumentiert der Anhang nur Auswertungen aus der Post-Befragung sowie der Follow-Up-Befragung. Hingegen sind offensichtlich keine Daten für den Zeitpunkt vor der Intervention verfügbar, welche einen Vorher-Nachher Vergleich erlauben würden. Ich konnte anhand des Berichtes nicht nachvollziehen, weshalb die Autoren der Auffassung sind, dass Veränderungen zwischen der Post-Befragung und der Followup Befragung ursächlich auf die Intervention zurückgeführt werden können. Daher wurde dieser Zielgrösse nicht weiter berücksichtigt.

Für die Zielgrössen *Stärken und Schwächen* sowie *Lebensqualität* hingegen enthalten die Berichte für alle Messzeitpunkte und für alle relevanten Teilbereiche die Mittelwerte, die Standardabweichungen und die Anzahl der Beobachtungen. Diese Informationen reichen aus, um zu berechnen, a) ob sich die Kontrollgruppe oder die Interventionsgruppe besser entwickelt hat, b) wie hoch die standardisierten Veränderungen (Effektstärken) innerhalb jeder Gruppe sind, c) ob sich die Veränderung in einer Interventionsgruppe statistisch bedeutsam von jener in der Kontrollgruppe unterscheidet.

Insbesondere kann man die folgenden Grössen berechnen:[6]

[6] Zur Berechnung der Effektstärken Cohen's d sowie der assoziierten Standardfehler und Konfidenzintervalle wurden EXCEL-Vorlagen des „Curriculum, Evaluation and Management Centre" der University of Durham verwendet. Die Tabelle sowie alle statistischen Erläuterungen

Die standardisierte Veränderung zwischen Prä-Messung und Follow-up Messung gemessen als Intra-Gruppen Effektstärke Cohen's d.

$$\text{Cohen's d} = \frac{M_{Pre} - M_{Follow-up}}{\sqrt{(SD^2_{Pre} + SD^2_{Follow-up})/2}} \qquad \textit{Gleichung 1}$$

Diese Effektstärken zeigen innerhalb einer Behandlungsbedingung die standardisierte Verbesserung zwischen der Vorher und der Follow-Up Messung. In Einklang mit der üblichen Praxis werden im folgenden alle standardisierten Effekte so gezeigt, dass ein positiver Effekt (d.h. > 0) eine Wirkung in die erwünschte Richtung bedeutet. Die im Anhang gezeigten Effekte sind so genannte Hedges korrigierte Effektstärken, allerdings sind die Unterschiede zu unkorrigierten Effekten minim.

Um die Standardfehler der Effektstärken zu berechnen wird die von Hedges und und Olkin (1985) vorgeschlagene Formel verwendet:

$$S.E_{\text{Cohen's d}} = \sqrt{\frac{N_{pre} + N_{follow-up}}{N_{pre} \times N_{follow-up}} + \frac{d^2}{2(N_{pre} + N_{follow-up})}} \qquad \textit{Gleichung 2}$$

Schliesslich sind noch die Inter-Gruppen Effekte als Differenz zwischen den Intra-Gruppen Effekte zu berechnen.

$$ES_{\text{between groups}} = ES_{\text{within(Treatment)}} - ES_{\text{within(control)}} \qquad \textit{Gleichung 3}$$

Sie zeigen, um wie viel besser oder schlechter sich die Behandlungsgruppe im Vergleich zur Kontrollgruppe entwickelt. Die Differenz wird üblicherweise so berechnet, dass positive Werte einem erwünschten Effekt in der Behandlungsgruppe entsprechen.

Um abzuschätzen, ob sich eine Interventionsgruppe von der Kontrollgruppe unterscheidet, wurde jeweils untersucht, ob sich die Vertrauensintervalle der geschätzten Effektstärken in Kontrollgruppe und Behandlungsgruppe überlappen. Manche Forschende glauben fälschlicherweise, dass hierzu auf eine Überlappung des 95% Vertrauensintervalls geprüft werden muss. Wie allerdings Payton et al. (2003) zeigen, entspricht eine Nicht-überlappung des 95% Vertrauensintervalls einem sehr viel höheren Signifikanztest als $p < 0.05$. Um einen Unterschied zweier annähernd normalverteilter Mittelwerte auf einem 5-Prozent Niveau statistisch abzusichern, genügt jedoch der Nachweis, dass sich die 84% Vertrauensintervalle nicht überlappen (Payton et al., 2003).

Ergebnisse

Die Tabellen mit allen Daten und Auswertungen sind in Anhang 1 dieses Dokumentes angeführt. Alle Mittelwerte, Standardabweichungen und Ns sind direkt den erwähnten Forschungsberichten entnommen. Die

finden sich unter http://www.cemcentre.org/renderpage.asp?linkID=30325017. Sie erlauben es dem Lesenden, alle Effekte unabhängig von der hier vorgelegten Argumentation nochmals nachzurechnen.

Richtigkeit der Daten, welche für die Berechnungen verwendet wurden, wurde mir von den Autoren des Berichtes bestätigt. Alle Berechnungen wurden unabhängig von weiteren Fachpersonen überprüft.

Die nachfolgende Tabelle 1 enthält eine Übersicht der Befunde aufgrund der Auswertungen im Anhang. Sie zeigt die Analysen für jeden Vergleich einer Interventionsgruppe mit der Kontrollgruppe, der aufgrund der vorliegenden Daten in den Bereichen *Stärken und Schwächen* sowie *Lebensqualität* möglich war.

Spalte 3 zeigt die Effektstärke *Cohen's d*. Dieses Mass ist eine standardisierte Grösse, welche Vergleiche der erzielten Effekte zwischen Studien und unterschiedlichen Populationen erlaubt. Gemäss allgemeinen Konventionen gelten Effekte von etwa d=0.2 als „kleine" Effekte und Effekte von etwa 0.5 als „mittlere" Effekte. Spalte 4 zeigt, ob sich die Kontrollgruppe (KG) oder die Interventionsgruppe (TG) zwischen Prä-Messung und Follow-Up Messung besser entwickelt haben. Spalte 5 zeigt, ob dieser Unterschied als statistisch mit einer Irrtumswahrscheinlichkeit von $p < 0.05$ als abgesichert gelten kann. Die Ergebnisse lassen sich wie folgt zusammenfassen.

- Insgesamt konnten aufgrund der verfügbaren Daten 27 Vergleiche (jeweils drei Vergleiche für jede der neun Teilbereiche) durchgeführt werden.

- Sieht man zunächst von der statistischen Signifikanz ab, dann sieht man, dass von insgesamt 27 Vergleichen zwischen Interventions- und Kontrollgruppe 15 Veränderungen zu Gunsten der Kontrollgruppe ausfallen. Insgesamt 9 Veränderungen fallen zu Gunsten einer der Interventionsgruppen aus. In zwei Fällen sind die Effektstärken in beiden Gruppen identisch.

- Der durchschnittliche Effekt über alle 27 Vergleiche hinweg beträgt d = -0.04, was einer leichten Begünstigung der Kontrollgruppe entspricht.

- Fast alle Effekte sind sehr klein und können statistisch nicht abgesichert werden. Nur bei zwei Vergleichen kann die Null-Hypothese, dass keine Unterschiede bestehen, zurückgewiesen werden. Bei insgesamt 27 geprüften Wirkungen entspricht dies in etwa der Zahl von signifikanten Effekten, die zufällig zu erwarten sind. Ein Effekt begünstigt die Kontrollgruppe, ein Effekt begünstigt eine der Interventionsgruppen.

- Beschränkt man die Analyse auf Effekte von d > +/- 0.10, dann ergeben sich acht Effekte zu Gunsten der Kontrollgruppe und zwei Effekte zu Gunsten einer der Interventionsgruppen.

Insgesamt führen diese Befunde zum Schluss, dass in den geprüften Zielgrössen keine systematischen Unterschiede zwischen Interventionsgruppen und Kontrollgruppe bestehen. Der isolierte positive Effekt bei den Schwächen der Kinder nach Elternurteil in der Triple P Gruppe kann keinesfalls ausreichen, um die überschwänglich positive Beurteilung durch die Studienautoren zu begründen.

Tabelle 1 *Effekte der Interventionen bei ESSKI, Zusammenfassung*

(1) Geprüfte Teildimension

(2) Vergleichsgruppe zur Kontrollgruppe

(3) Inter-Gruppen Effektstärke Cohen's d = Intragruppen ES_{TG} – Intragruppen ES_{KG}. Positive Werte bedeuten bessere Entwicklung in der Treatment-Gruppe.

(4) Bessere Entwicklung in Kontrollgruppe (KG) oder Interventionsgruppe (TG) aufgrund des Vorzeichens von (3).

(5) Statistische Signifikanz des Effekts in (3) auf 95%-Niveau, basierend auf Konfidenzintervallen der Intra-Gruppen ES.

(1)	(2)	(3) Effektstärke Cohen's d	(4) Bessere Entwicklung in KG oder TG		(5) Unterschied statistisch signifikant
Stärken der Kinder. Lehrerurteil***	Fit fürs Leben	-0.20		KG	nein
	Triple P	-0.11		KG	nein
	Kombi	-0.34		KG	Ja, negativ
Schwächen der Kinder, Lehrerurteil	Fit fürs Leben	-0.15		KG	nein
	Triple P	-0.01		KG	nein
	Kombi	+0.02	TG		nein
Stärken der Kinder, Elternurteil	Fit fürs Leben	-0.02		KG	nein
	Triple P	-0.02		KG	nein
	Kombi	+0.10	TG		nein
Schwächen der Kinder, Elternurteil	Fit fürs Leben	+0.03	TG		nein
	Triple P	+ 0.38	TG		Ja, positiv
	Kombi	+0.16	TG		nein
Beziehung Kind-Eltern	Fit fürs Leben	+0.01	TG		nein
	Triple P	+0.00	=	=	nein
	Kombi	+0.01	TG		nein
Selbstwertgefühl Kinder***	Fit fürs Leben	+0.03	TG		nein
	Triple P	+0.07	TG		nein
	Kombi	-0.10		KG	nein
Spass an der Schule	Fit fürs Leben	-0.09		KG	nein
	Triple P	-0.15		KG	nein
	Kombi	-0.21		KG	nein
Spass und Lachen***	Fit fürs Leben	-0.05		KG	nein
	Triple P	-0.09		KG	nein
	Kombi	+0.04	TG		nein
Beziehung zu Freunden***	Fit fürs Leben	-0.21		KG	nein
	Triple P	0.00	=	=	nein
	Kombi	-0.15		KG	nein

*** Diese Bereiche wurden aus unbekannten Gründen im deutschsprachigen ESSKI-Bericht nicht aufgeführt. Die entsprechenden Daten wurden dem englischsprachigen Bericht entnommen. In der Tendenz begünstigen die im deutschsprachigen Bericht weggelassenen Teilbereiche eher die Kontrollgruppe.

Mögliche Gründe für die unterschiedlichen Folgerungen

Insgesamt ergibt sich der Schluss, dass mit anerkannten Verfahren der Sekundäranalyse von publizierten Interventionsstudien nur ein signifikanter positiver Effekt der realisierten Interventionen nachgewiesen werden kann, dem ein negativer Effekt gegenübersteht. Daher stellt sich die Frage, weshalb das ESSKI-Team in den Projektberichten zu einer völlig anderen Einschätzung der Daten gekommen ist und im Schlussbericht davon ausgeht, dass *alle Ergebnisse signifikant* gewesen seien (Schönenberger, Schmid et al., 2006: 20).

Leider geben die vorliegenden Berichte keine genaue Auskunft, wie die Daten vom ESSKITeam analysiert wurden und genau auf welche Auswertungen sich die Aussagen abstützen. Der englische Bericht enthält allerdings einen Abschnitt zum Vorgehen bei der Datenanalyse. Ich zitiere den Abschnitt vollständig: „Data were analysed using SPSS for Windows. Absolute and relative frequency of baseline characteristics were given by group and compared

with chi-square tests. The comparison between pre-, post-test and follow-up values were separately reported for teachers, parents and students. Means, standard deviations and the number of subjects were given for scales, and absolute as well as relative frequencies were given for indicators. For teachers, effects of group and time of measurement were tested applying a two way analysis of variance with one repeated factor (ANOVAR). Our main hypothesis, that Condition 3 (Fit fürs Leben + Triple P) is the most effective in terms of health promotion for participants can be tested thru the interaction between group and time. The same analysis was also applied on the individual level for parents as well as for students, using the general linear modeling approach. In order to control for the clustering in school classes, however, we also report results from a mixed model analysis where we tested for significant variation of the intercepts. The test represents differences between school classes and the analysis controls for possible differences in the primary sampling unit of school classes." (Schmid et al., 2007: 9)

Das Vorgehen der Autoren wird besser verständlich, wenn man sich die Outputs im Anhang des englischen oder deutschen Berichtes anschaut. Die Ergebnisse lassen erkennen, dass für jedes Outcome varianzanalytische Methoden eingesetzt und zwei Gleichungen geschätzt wurden.

Gleichung 1 Outcome = GROUP + TIME + (GROUP*TIME)

Gleichung 2 Outcome = GROUP + TIME + (GROUP*TIME) + Intercept

Wie man sieht, wurden bei jeder Gleichung zwei Haupteffekte sowie ein Interaktionsterm geschätzt. Aus der obigen Beschreibung der durchgeführten Analysen konnte ich nicht nachvollziehen, welche Komponente der Gleichungen die Autoren selbst als Evidenz für die Wirksamkeit der Interventionen ansehen. Die folgenden Erläuterungen versuchen daher zu klären, was durch die realisierten Analysen tatsächlich geprüft wurde.

GROUP misst den Haupteffekt nach Gruppen im Durchschnitt aller Messungen. Dieser Koeffizient zeigt an, ob sich die Mittelwerte der vier Gruppen in irgendeiner Weise voneinander unterscheiden. In den Berichten werden keine Ex-post Tests gezeigt, mit denen untersucht werden kann, welche Gruppen sich voneinander in welche Richtung unterscheiden. Allerdings sind zeitinvariante Gruppenunterschiede für eine Beurteilung der Interventionseffekte nicht von Belang.

TIME misst, ob insgesamt, d.h. im Durchschnitt aller vier Gruppen, zwischen den drei Zeitpunkten (Prä, Post, Follow-Up) Unterschiede bestehen. Da offensichtlich kein linearer Effekt von Zeit spezifiziert wurde, kann ohne zusätzliche Analysen nicht näher bestimmt werden, in welcher Weise sich die Mittelwerte zwischen den drei Messzeitpunkten unterscheiden. Ein signifikanter Effekt kann heissen, dass die Mittelwerte insgesamt angestiegen sind, dass sie gesunken sind, oder dass sie ein U-förmiges Muster zeigen. Zeiteffekte – d.h. ein gesamthaftes Ansteigen oder Absinken der Mittelwerte in allen vier Gruppen - sind für eine Beurteilung der Wirkung der Interventionen ebenfalls nicht von Belang.

Zur Bestimmung von Interventionseffekten ist einzig die GROUP*TIME-Interaktion von Interesse. Allerdings besagt ein signifikanter Interaktionsterm lediglich, dass sich zwei oder mehr Gruppen über die Zeit voneinander verschieden unterscheiden. Er sagt hingegen nichts darüber aus, wo, d.h. zwischen welchen Gruppen, Unterschiede vorliegen und in welcher Richtung diese verlaufen. Es kann vorliegend also durchaus sein, dass sich eine signifikante GROUP*TIME_Interaktion aus einem unterschiedlichen Verlauf zwischen der Triple P und Fit-fürs-Leben Gruppe ergibt, während die Kontroll- und Kombigruppe eine mittlere Position einnehmen und sich weder von der einen noch von der anderen Gruppe unterscheiden. Es kann ebenso vorkommen, dass sich die Kontrollgruppe von allen anderen drei Gruppen unterscheidet, weil sie einen günstigeren Verlauf aufweist als die Behandlungsgruppen. In beiden Beispielen ergibt sich ein signifikanter Effekt, der aber keineswegs eine Wirkung der Behandlungen nachweisen kann.

Zur Erbringung des Wirksamkeitsnachweises bedarf es daher zweier weiterer Analyseschritte, wobei nur zwei Messzeitpunkte verglichen werden sollten (Prä- vs. Postmessung oder Prä- vs. Follow-up-Messung, ansonsten könnten auch irrelevante Gruppenunterschiede zwischen Postund Follow-up-Messung die Quelle signifikanter Effekte sein): Zunächst gilt es nach der Identifizierung eines signifikanten GROUP*TIME-Interaktionseffekts dieser

genauer zu lokalisieren, d.h. zu bestimmen, welche Gruppen sich voneinander unterscheiden. Dazu steht eine ganze Palette so genannter Post-hoc-Tests zur Verfügung, wobei diese allerdings noch nichts über die Richtung von Gruppenunterschieden aussagen. Stellt sich nun heraus, dass sich die Kontrollgruppe von einer oder mehreren Versuchsgruppen unterscheidet, gilt es daher in einem letzten Schritt anhand eines Mittelwertevergleichs zu eruieren, ob die Kontrollgruppe sich tatsächlich schlechter entwickelt hat als die sich signifikant unterscheidende Versuchsgruppe. Erst jetzt kann von einer signifikanten Programmwirkung die Rede sein.

INTERCEPT misst Abweichungen der einzelnen Schulhausmittelwerte vom Gesamtmittelwert und ist nur als Kontrollgrösse zur präziseren Schätzung der interessierenden Effekte von Bedeutung.

Folgerungen

Die vorliegende Reanalyse führt zum Schluss, dass die im Rahmen von ESSKI realisierten Interventionen keine nachweisbaren Effekte in jenen zwei Bereichen hatten, die von der Autorengruppen selbst in den Mittelpunkt gerückt werden. *Aufgrund der berechneten Effektstärken lassen sich weder für Stärken und Schwächen der Kinder im Elternurteil, noch für Stärken und Schwächen im Lehrerurteil, noch für die Teildimensionen von Lebensqualität systematisch positive Effekte nachweisen. Für den dritten Bereich, Tabakkonsum, fehlen Daten aus der ersten Befragung vor der Intervention.* Solche Daten wären zwingend notwendig, um allfällige Wirkungen in diesem Bereich zu bestimmen.

Selbstverständlich hat die hier vorgelegte Reanalyse Grenzen. Beispielsweise habe ich keine weiteren möglichen Zielgrössen betrachtet, welche für eine Beurteilung der Studie von Belang sein könnten. Ich habe also nicht untersucht, ob die Intervention allenfalls positive Effekte auf Bereiche hatte, welche in der Medienmitteilung kaum oder gar nicht erwähnt wurden. Ebenso habe ich aus den oben geschilderten Gründen den Vergleich auf die Follow-Up Messung fünf Monate nach der Intervention beschränkt. Es ist möglich, dass kurzfristig positive Effekte erzielt wurden. Man sollte auch beachten, dass die vier Gruppen zu Beginn der Studie nur bedingt äquivalent waren. Entsprechend wäre es für eine detaillierte Analyse angebracht, Anfangswerte zu t0 als Kovariate einzuschliessen. Schliesslich wurden bei der vorliegenden Studie die Teilnehmenden nicht als Individuen sondern in Gruppen (d.h. hier in Klassen) den Experimentalbedingungen zugeteilt. Solche klumpenrandomisierte Designs führen zu einem weiteren Verlust an statistischer Power (Campbell, Mollison, Steen, Grimshaw, & Eccles, 2000; Donner & Klar, 2004). Ihre Berücksichtigung würde eine Mehrebenenanalyse bedingen, die aufgrund der publizierten Daten nicht möglich ist. Es könnte aber sein, dass anspruchsvollere statistische Modelle zu etwas anderen Ergebnissen führen würden. Allerdings sollten Befunde aus komplexen statistischen Modellen, die zu Ergebnissen füh ren, welche einer direkten Analyse von Effektstärken widersprechen, besonders sorgfältig geprüft werden.

Festzuhalten bleibt, dass für eine breitere Öffentlichkeit und die Praxis Wirksamkeit letztlich bedeutet, dass sich die Zielgrössen in der Treatmentkondition besser entwickeln als in der Kontrollgruppe. Das ist bei der vorliegenden Studie, wie oben gezeigt, bei mehr als 50 Prozent der Zielvariablen (15 von 27 Vergleichen) nicht der Fall. Es ist durchaus denkbar, dass Forschende trotzdem zum Schluss kommen, die Interventionen seien wirksam gewesen. Allerdings sollte in einem solchen Fall ganz besonders sorgfältig darauf geachtet werden, das methodische Vorgehen detailliert zu dokumentieren.[7]

Die Ergebnisse der vorliegenden Studie legen verschiedene allgemeine Schlussfolgerungen nahe. Insbesondere ist es in Deutschland und der Schweiz während der letzten zehn Jahre zu einer deutlichen Zunahme von randomisierten Feldstudien gekommen, welche zum Ziel haben, wirksame Programme zur Prävention von

[7] Ich habe den Autoren der ESSKI-Studie über die hier vorgebrachten Kritikpunkte vor der Publikation dieses Berichtes ins Bild gesetzt um mit ihnen abzuklären, ob sie die vorgebrachten Argumente nachvollziehen und entsprechend den Forschungsbericht modifizieren möchten. Nach Auffassung der Autoren sind die publizierten Ergebnisse jedoch methodisch korrekt zustande gekommen. Ebenso geben nach Auffassung der Autoren die Schlussfolgerungen die erhaltenen Ergebnisse korrekt wieder. Es ist daher festzuhalten, dass wir in dieser Hinsicht unterschiedliche Auffassungen vertreten.

Problemverhalten und zur Förderung einer gesunden Entwicklung für die Praxis zu identifizieren. Das ist eine positive Entwicklung, welche hoffentlich zu einer wirksameren Prävention in der Zukunft führen wird. Allerdings fehlen gegenwärtig im gesamten Bereich von psycho-sozialen Präventionsexperimenten wirksame Mechanismen der Qualitätskontrolle, welche etwa Richtlinien über die adäquate Präsentation von Ergebnissen nicht nur in wissenschaftlichen Zeitschriften, sondern vor allem auch gegenüber der Öffentlichkeit enthalten müssten. Es wäre höchst wünschenswert, solche Standards für die Durchführung und Präsentation von Evaluationsstudien im deutschsprachigen Raum einzuführen. Solche Standards sind mindestens teilweise schon verfügbar.

Farrington (2003) hat beispielsweise ein umfassende Liste von Qualitätsstandards für die Evaluationsforschung vorgeschlagen, welche gut verständlich ist und beispielsweise auch von Praktikern, Politikern und Medienschaffenden genutzt werden kann. Sie enthält beispielsweise eine Liste von Elementen, die in jedem Evaluationsbericht als Minimalanforderung vorhanden sein müssen. Hierzu gehören etwa gut verständlich dargestellte Effektstärken für alle relevanten Zielgrössen oder eine Darlegung der möglichen Interessenkonflikte der Forschenden. Farrington (2003: 55) schlägt auch vor, dass Fachvereinigungen und Geldgeber von Forschungsprojekten gemeinsam Kriterien aufstellen, wie Berichte über Wirksamkeitsstudien zu erstellen sind und welche Informationen enthalten sein müssen. Solche Standards würden einen grossen Beitrag liefern, um das öffentliche Vertrauen in die Ergebnisse von sozialwissenschaftlicher Evaluationsforschung zu stärken.

ANHANG 1 Datenauswertung

Hinweis: Alle Mittelwerte, Standardabweichungen und Ns wurden den Anhängen der Projektberichte aus der ESSKI-Studie entnommen. Zur Berechnung der Effektstärken und Vertrauensintervalle siehe die Formeln im Text oben. Die Vertrauensintervalle umfassen 84 Prozent der erwarteten Verteilung der Effektstärken. Ein Nicht—Überlappen der verglichenen Vertrauensintervalle entsprich einem Test auf eine statistische Signifikanz von $p < 0.05$.

A) Stärken und Schwächen der Kinder

Tabelle 1 Stärken (soziale Kompetenzen) der Kinder nach Lehrereinschätzung

	Mittelwert (Standardabweichung)			Effektstärke	Vertrauensintervall	
	Vorher	Follow-Up	Differenz	Cohens d	Untere	obere
Kontrollgruppe (N = 303)	2.41 (0.40)	2.52 (0.36)	+0.11	**+0.29**	+0.17	+0.40
„Fit fürs Leben" (N = 317)	2.40 (0.44)	2.44 (0.41)	+0.04	**+0.09**	-0.02	+0.21
Triple P Gruppe (N = 357)	2.40 (0.40)	2.47 (0.37)	+0.07	**+0.18**	+0.08	+0.29
Kombinierte G. (N = 265)	2.49 (0.38)	2.47 (0.37)	-0.02	**-0.05***	-0.18	+0.07

Folgerungen
Nach Einschätzung der Lehrpersonen erfolgte die grösste Zunahme der sozialen Kompetenzen in der Kontrollgruppe. Der Effekt für die kombinierte Gruppe liegt ausserhalb des Konfidenzintervalls in der Kontrollgruppe. Die kombinierte Gruppe hat sich statistisch signifikant schlechter entwickelt als die Kontrollgruppe.

Tabelle 2 Schwächen (Problemverhalten) der Kinder nach Lehrereinschätzung

	Mittelwert (Standardabweichung)			Effektstärke	Vertrauensintervall	
	Vorher	Follow-Up	Differenz	Cohens d	Untere	obere
Kontrollgruppe (N = 303)	1.42 (0.34)	1.35 (0.31)	-0.07	**+0.21**	+0.10	+0.33
„Fit fürs Leben" (N = 317)	1.42 (0.33)	1.40 (0.33)	-0.02	**+0.06**	-0.05	+0.17
Triple P Gruppe (N = 357)	1.37 (0.30)	1.31 (0.29)	-0.06	**+0.20**	+0.10	+0.31
Kombinierte G. (N = 265)	1.38 (0.32)	1.31 (0.30)	-0.07	**+0.23**	-0.11	+0.35

Folgerungen
Bei allen Gruppen haben die Schwächen der Kinder abgenommen. Es gibt keine statistisch signifikanten Unterschiede zwischen den Gruppen.

Tabelle 3 Stärken (Soziale Kompetenzen) der Kinder nach Elterneinschätzung

	Mittelwert (Standardabweichung)		Differenz	Effektstärke Cohens d	Vertrauensintervall	
	Vorher	Follow-Up			Untere	obere
Kontrollgruppe (N = 140)	2.51 (0.30)	2.56 (0.28)	+0.05	**+0.17**	-0.00	+0.34
„Fit fürs Leben" (N = 166)	2.44 (0.33)	2.49 (0.34)	+0.05	**+0.15**	-0.01	+0.30
Triple P Gruppe (N = 126)	2.54 (0.28)	2.58 (0.25)	+0.04	**+0.15**	-0.03	+0.33
Kombinierte G. (N = 127)	2.46 (0.29)	2.54 (0.29)	+0.08	**+0.27**	-0.10	+0.45

Folgerungen
Bei allen Gruppen haben die Stärken der Kinder nach Elternurteil zugenommen. Keine der Interventionsgruppen unterscheidet sich signifikant von der Kontrollgruppe.

Tabelle 4 Schwächen (Problemverhalten) der Kinder nach Elterneinschätzung

	Mittelwert (Standardabweichung)		Differenz	Effektstärke Cohens d	Vertrauensintervall	
	Vorher	Follow-Up			Untere	obere
Kontrollgruppe (N = 140)	1.40 (0.29)	1.38 (0.32)	-0.02	**+0.07**	-0.10	+0.23
„Fit fürs Leben" (N = 166)	1.45 (0.30)	1.42 (0.30)	-0.03	**+0.10**	-0.05	+0.25
Triple P Gruppe (N = 126)	1.48 (0.30)	1.36 (0.23)	-0.12	**+0.45***	+0.27	+0.63
Kombinierte G. (N = 127)	1.45 (0.30)	1.38 (0.30)	-0.07	**+0.23**	-0.06	+0.41

Folgerungen
Bei allen Gruppen haben die Schwächen der Kinder nach Elternurteil abgenommen. In der Triple P Gruppe ist die Verbesserung statistisch signifikant besser als in der Kontrollgruppe.

B) Lebensqualität der Kinder

Tabelle 5 Beziehung Kind-Eltern nach Einschätzung des Kindes

	Mittelwert (Standardabweichung)			Effektstärke	Vertrauensintervall für d	
	Vorher	Follow-Up	Differenz	Cohen's d	Untere	obere
Kontrollgruppe (N = 254)	2.64 (0.56)	2.69 (0.56)	+0.05	**+0.09**	-0.04	+0.21
„Fit fürs Leben" (N = 244)	2.71 (0.54)	2.76 (0.47)	+0.05	**+0.10**	-0.03	+0.23
Triple P Gruppe (N = 241)	2.67 (0.55)	2.72 (0.54)	+0.05	**+0.09**	-0.04	+0.22
Kombinierte G. (N = 165)	2.65 (0.56)	2.70 (0.48)	+0.05	**+0.10**	-0.06	+0.25

Folgerungen
Es gibt keinerlei Unterschiede zwischen den Gruppen bezüglich der Veränderung der Beziehung Eltern-Kind.

Tabelle 6 Selbstwertgefühl nach Einschätzung des Kindes

	Mittelwert (Standardabweichung)			Effektstärke	Vertrauensintervall für d	
	Vorher	Follow-Up	Differenz	Cohen's d	Untere	obere
Kontrollgruppe (N = 253)	2.51 (0.58)	2.58 (0.58)	+0.07	**+0.12**	-0.00	+0.25
„Fit fürs Leben" (N = 247)	2.55 (0.56)	2.63 (0.53)	+0.08	**+0.15**	-0.02	+0.27
Triple P Gruppe (N = 243)	2.47 (0.55)	2.59 (0.57)	+0.12	**+0.21**	-0.09	+0.34
Kombinierte G. (N = 167)	2.60 (0.52)	2.61 (0.51)	+0.01	**+0.02**	-0.13	+0.17

Folgerungen
Es gibt keine statistisch abgesicherten Unterschiede zwischen den Gruppen bezüglich der Veränderung des Selbstwertgefühls

Tabelle 7 Spass an der Schule nach Einschätzung des Kindes

	Mittelwert (Standardabweichung)		Differenz	Effektstärke Cohen's d	Vertrauensintervall für d	
	Vorher	Follow-Up			Untere	obere
Kontrollgruppe (N = 254)	2.45 (0.65)	2.53 (0.59)	+0.08	**+0.13**	-0.00	+0.25
„Fit fürs Leben" (N = 244)	2.58 (0.61)	2.59 (0.58)	+0.01	**+0.02**	-0.11	+0.14
Triple P Gruppe** (N = 221)	2.58 (0.60)	2.57 (0.59)	-0.01	**-0.02**	-0.15	+0.12
Kombinierte G. (N = 165)	2.73 (0.49)	2.69 (0.53)	-0.04	**-0.08**	-0.23	+0.08

** In dieser Tabelle liegt für die Triple P-Gruppe das N mit 221 deutlich tiefer als bei allen anderen Teildimensionen (zwischen N = 239 und N = 243).

Folgerungen
Es gibt keine statistisch nachweisbaren Unterschiede zwischen den Gruppen.

Tabelle 8 Spass und Lachen nach Einschätzung des Kindes

	Mittelwert (Standardabweichung)		Differenz	Effektstärke Cohen's d	Vertrauensintervall für d	
	Vorher	Follow-Up			Untere	obere
Kontrollgruppe (N = 253)	2.59 (0.58)	2.58 (0.58)	-0.01	**-0.02**	-0.11	+0.09
„Fit fürs Leben" (N = 246)	2.67 (0.55)	2.63 (0.53)	-0.04	**-0.07**	-0.05	+0.20
Triple P Gruppe (N = 242)	2.65 (0.55)	2.59 (0.57)	-0.06	**-0.11**	-0.02	+0.23
Kombinierte G. (N = 167)	2.66 (0.52)	2.67 (0.51)	+0.01	**+0.02**	-0.17	+0.13

Folgerungen
Es gibt keine Unterschiede zwischen den Gruppen.

Tabelle 9 Gute Beziehungen zu Freunden nach Einschätzung des Kindes

	Mittelwert (Standardabweichung)		Differenz	Effektstärke Cohen¨s d	Vertrauensintervall für d	
	Vorher	Follow-Up			Untere	obere
Kontrollgruppe (N = 251)	2.71 (0.56)	2.78 (0.48)	+0.07	**+0.13**	-0.01	+0.26
„Fit fürs Leben" (N = 242)	2.81 (0.46)	2.77 (0.49)	-0.04	**-0.08**	-0.21	+0.04
Triple P Gruppe (N = 239)	2.69 (0.55)	2.76 (0.49)	+0.07	**+0.13**	-0.01	+0.26
Kombinierte G. (N = 165)	2.78 (0.49)	2.77 (0.49)	-0.01	**-0.02**	-0.17	+0.13

Folgerungen
Die „Fit fürs Leben" Gruppe hat sich hinsichtlich der guten Beziehungen zu Freunden am schlechtesten entwickelt. Der Unterschied ist allerdings statistisch knapp nicht signifikant

Literatur

Bullinger, M., von Mackensen, S., & Kirchberger, I. (1994). KINDL - Ein Fragebogen zur Erfassung der gesundheitsbezogenen Lebensqualität von Kindern. *Zeitschrift für Gesundheitspsychologie, 1*, 64-77.

Burow, F., Asshauer, M., & Hanewinkel, R. (1998). *Fit und stark fürs Leben. 1. und 2. Schuljahr. Persönlichkeitsförderung zur Prävention von Aggression, Rauchen und Sucht*. Leipzig: Ernst Klett Grundschulverlag.

Campbell, M. K., Mollison, J., Steen, N., Grimshaw, J. M., & Eccles, M. (2000). Analysis of cluster randomized trials in primary care: a practical approach. *Fam. Pract., 17*(2), 192-196.

Donner, A., & Klar, N. (2004). Pitfalls of and Controversies in Cluster Randomization Trials. *Am J Public Health, 94*(3), 416-422.

Eisner, M., Ribeaud, D., & Bittel, S. (2006). *Prävention von Jugendgewalt: Wege zu einer evidenzbasierten Gewaltprävention*. Bern: Eidgenössische Ausländerkommission.

Eltern und Schule stärken Kinder. (2007). ESSKI-Projekt macht Kinder stark (Media release, 18 January 2007, available at www.esski.ch, last accessed on 16 March 2007).

Farrington, D. P. (2003). Methodological Quality Standards for Evaluation Research. *Annals of the American Academy of Political and Social Sciences, 587*, 49-68.

Farrington, D. P., & Welsh, B. (2003). Family-based Prevention of Offending: A Meta-analysis. *Australian and New Zealand Journal of Criminology, 36*(2), 127-151.

Goodman, R. (2001). Psychometric Properties of the Strengths and Difficulties Questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry, 40*(11), 1337-1345.

Hedges, L., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. New York: Academic Press.

Littell, J. (2005). Lessons from a systematic review of effects of multisystemic therapy. *Children and Youth Services Review, 27*(4), 445-463.

Lösel, F., & Beelmann, A. (2003). Effects of Child Skills Training in Preventing Antisocial Behavior: A Systematic Review of Randomized Evaluations. *The ANNALS of the American Academy of Political and Social Science, 587*(1), 84-109.

Payton, M. E., Greenstone, M. H., & Schenker, N. (2003). Overlapping Confidence Intervals or Standard Error Intervals: What do They Mean in Terms of Statistical Significance. *Journal of Insect Science, 3*(34), 34.

Ravens-Sieberer, U. (1998). Assessing health-related quality of life in chronically ill children with the German KINDL: first psychometric and content analytical results. *Quality of Life research, 7*(5), 399-407.

Reyno, S. M., & McGrath, P. J. (2006). Predictors of parent training efficacy for child externalizing behavior problems - a meta-analytic review. *Journal of Child Psychology and Psychiatry, 47*(1), 99-111.

Sanders, M. R., Markie-Dadds, C., & Turner, K. T. (2003). Theoretical, Scientific and Clinical Foundations of the Triple P Positive Parenting Program: A Population Approach to the Promotion of Parenting Competence. *Parenting Research and Practice Monograph, 1*, 1-21.

Schmid, H., Anliker, S., Bodenmann, G., Cina, A., Fäh, B., Kern, W., et al. (2007). *Empowerment in family and schools (eifas): A randomized controlled trial (unpublished report)*.

Schönenberger, M., Lattmann, U. P., Fäh, B., Schmid, H., Bodenmann, G., Cina, A., et al. (2006). Eltern und Schule stärken Kinder" (ESSKI); Konzept eines mehrdimensionalen Forschungs- und Entwicklungsprojektes im Bereich psychosoziale Gesundheit in Schule und Elternhaus. *Abhängigkeiten* (3).

Schönenberger, M., Schmid, H., Fäh, B., Bodenmann, G., Lattmann, U. P., Cina, A., et al. (2006). *Projektbericht "Eltern und Schule stärken Kinder" (ESSKI); Ein Projekt zur Förderung der Gesundheit bei Lehrpersonen, Kindern und Eltern und zur Prävention von Stress, Aggression und Sucht - Ergebnisse eines mehrdimensionalen*

Forschungs- und Entwicklungsprojekts im Bereich psychosoziale Gesundheit in Schule und Elternhaus. (report available at http://esski.ch/html/download.htm).

Shadish, W. R., Cook, T. D., & Campbell, D. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.

St Pierre, T. L., Osgood, D. W., Mincemoyer, C. C., Kaltreider, D. L., & Kauh, T. J. (2006). Results of an Independent Evaluation of Project ALERT Delivered in Schools by Cooperative Extension. *Prevention Science, 6*(4), 305-317.

Wilson, S. J., & Lipsey, M. W. (2006). The Effects of School-based Social Information Processing Interventions on Aggressive Behavior, Part I: Universal Programs (Campbell Collaboration Systematic Review, http://www.campbellcollaboration.org/doc-pdf/wilson_socinfoprocuniv_review.pdf).

Wilson, S. J., Lipsey, M. W., & Derzon, J. H. (2003). The Effects of School-Based Intervention Programs on Aggressive Behavior: A Meta-Analysis. *Journal of Consulting and Clinical Psychology, 71*(1), 136-149.